Inf Retrieval (2013) 16:267–305
DOI 10.1007/s10791-012-9206-z

CROWD SOURCING

Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems

Guido Zuccon · Teerapong Leelanupab · Stewart Whiting · Emine Yilmaz · Joemon M. Jose · Leif Azzopardi

Received: 20 May 2011 / Accepted: 26 June 2012 / Published online: 13 July 2012

Abstract In the field of information retrieval (IR), researchers and practitioners are often faced with a demand for valid approaches to evaluate the performance of retrieval systems. The Cranfield experiment paradigm has been dominant for the in-vitro evaluation of IR systems. Alternative to this paradigm, laboratory-based user studies have been widely used to evaluate interactive information retrieval (IIR) systems, and at the same time investigate users' information searching behaviours. Major drawbacks of laboratory-based user studies for evaluating IIR systems include the high monetary and temporal costs involved in setting up and running those experiments, the lack of heterogeneity amongst the user population and the limited scale of the experiments, which usually involve a relatively restricted set of users. In this paper, we propose an alternative experimental methodology to laboratory-based user studies. Our novel experimental methodology uses a crowdsourcing platform as a means of engaging study participants. Through crowdsourcing, our experimental methodology can capture user interactions and searching behaviours at a lower cost, with more data, and within a shorter period than traditional laboratory-based

G. Zuccon (✉)
Australian e-Health Research Centre, CSIRO, Brisbane, QLD, Australia
e-mail: guido.zuccon@csiro.au

T. Leelanupab
King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
e-mail: teerapong@it.kmitl.ac.th

S. Whiting · J. M. Jose · L. Azzopardi
School of Computing Science, University of Glasgow, Glasgow, UK
e-mail: Stewart.Whiting@glasgow.ac.uk

J. M. Jose
e-mail: Joemon.Jose@glasgow.ac.uk

L. Azzopardi
e-mail: Leif.Azzopardi@glasgow.ac.uk

E. Yilmaz
Microsoft Research, Cambridge, UK
e-mail: eminey@microsoft.com

user studies, and therefore can be used to assess the performances of IIR systems. In this article, we show the characteristic differences of our approach with respect to traditional IIR experimental and evaluation procedures. We also perform a use case study comparing crowdsourcing-based evaluation with laboratory-based evaluation of IIR systems, which can serve as a tutorial for setting up crowdsourcing-based IIR evaluations.

Keywords Crowdsourcing evaluation · Interactive IR evaluation · Interactions

1 Introduction

Traditional information retrieval (IR) evaluation methodology follows the para-digm put forward by the Cranfield experiments, which is based "on the abstraction of a test collection" (Voorhees 2005): a set of documents, a set of topics and a set of relevance judgements from topic experts. Users are thus abstracted from the evaluation paradigm. This methodology has been the corner stone of evaluation campaigns such as TREC (Voorhees and Harman 2005):[1] it is however not suited to the evaluation of interactive IR (IIR) systems. In IIR experiments, users and their interactions with retrieval systems are essential for both the evaluation of systems and for the understanding of users' search behaviours. Commonly, IIR evaluation is conducted through laboratory-based user studies, where users are invited to perform some pre-defined simulated information seeking tasks.

Crowdsourcing has been used as an inexpensive and often efficient way to conduct large-scale studies. The crowdsourcing paradigm has already been successfully used in IR for performing a number of tasks (Alonso and Mizzaro 2009; Alonso et al. 2008; McCreadie et al. 2011).

In this paper, we focus on the evaluation of IIR systems. We however depart from common IIR evaluation approaches, such as laboratory-based user studies, and propose to evaluate systems using crowdsourcing. We do not list a series of rules crowdsourcing-based IIR evaluation must attain. Instead, we outline some issues with this methodology of evaluation and suggest solutions we have found and experimented with. The approach we propose here is based on the methodology for capturing search interactions investigated by Zuccon et al. (2011a, b), who show that crowdsourcing can be effectively used to build up query-logs and interaction-logs.

In order to assess the validity of our proposal, we compare and contrast our crowdsourced-based paradigm with the common laboratory-based user study methodology. We not only analyse commonalities and differences, but also employ both approaches for evaluating the performances of two IIR systems, comparing the results obtained from each study. We show that crowdsourced-based IIR evaluation can be reliably used to evaluate IIR systems, and that the provided insights are comparable to those of a traditional laboratory-based user study. This suggests that crowdsourcing-based IIR evaluation can be performed alongside other traditional IIR evaluation strategies, or by itself, so as to inform IR researchers and practitioners on the performance of their IIR systems.

The paper develops as follows. In Sect. 2 we detail the proposal of a crowdsourcing-based evaluation paradigm for IIR systems and we examine its components. In Sect. 3 we compare and contrast the crowdsourcing-based approach with the common laboratory-based user study approach in the ambit of IIR evaluation. This comparison is continued in Sect. 4 where we present an empirical study comparing the two approaches "in the field",

[1] The TREC Interactive Track (e.g. see Over 1997, 2001) represents a notable exception.

i.e. when used to evaluate two IIR systems. The results of the empirical study are presented in Sect. 5, while in Sect. 6 we analyse the obtained results comparing the different settings. Section 7 summarises our findings. In Sect. 8 we revisit related works on the evaluation of IIR systems and on the use of crowdsourcing for relevance evaluation. Finally, we outline directions for future research in Sect. 9 and conclude the paper summarising the contributions of this article in Sect. 10.

2 Evaluating IIR systems using crowdsourcing

In the following we describe our experiment methodology based on crowdsourcing, which we show can be used to evaluate IIR systems. Some of the considerations we develop in the following are based on the tools provided by Amazon Mechanical Turk (AMT),[2] but can be extended and adapted to other web-based crowdsourcing platforms, such as CrowdFlower[3] for example.

The methodology we propose prescribes that workers are asked to perform self-contained information seeking tasks within a unit of work (also known as a Human Intelligence Task, or, HIT in AMT) advertised on the crowdsourcing platform.

Researchers can collect logs of workers' interactions with the IIR system as well as post-search information and statistics. Researchers can also capture information to characterise the user population, both before and while workers perform HITs. The final output of such processes consists of

(1) A rich log of the workers interactions with the IIR system;
(2) Qualitative data the workers provide to describe their search experience and interaction with the system; and
(3) A quantitive assessment of each worker's abilities.

Although this procedure might appear to be similar to laboratory-based IIR studies, a number of key factors affect important experimental aspects, thus effectively differentiating these two methodologies. Next, we examine these differences.

2.1 Characterise user population

In laboratory-based IIR studies, pre-experiment questionnaires and interviews are usually employed by researchers for acquiring demographical (e.g. sex, age, nationality, etc) and self-perceptual information (i.e. familiarity/confidence with tasks, tools, etc) about participants.

This method is requires modifications to be applicable to crowdsourced IIR experiments. If workers are asked to fill in questionnaires[4] within a unit of work (i.e. HIT), then they will have to enter the same information several times: as many as the number of HITs they perform. This problem can be overcome by requiring workers to pass a *qualification test*. By employing qualification tests, researchers can acquire background information about the users to characterise the user population. Furthermore, experimenters can exclude from the HITs those workers that do not meet pre-defined criteria for the

[2] http://www.mturk.com/.

[3] http://crowdflower.com/.

[4] We ignore the possibility of performing interviews of workers, given the remote and asymmetric nature of crowdsourcing.

experiment, e.g. workers that have no or too much prerequisite knowledge about the search topics. Thus, within our experiment methodology qualification tests can absolve two roles: characterise the user population and select good workers for the search tasks. Regarding the second issue, the work of Alonso and Baeza-Yates (2011) provides a number of suggestions and alternatives for filtering workers.

Once workers are characterised through a qualification test, they can be classified within groups on the basis of similar scores. The intuition is that users from similar backgrounds and with similar skills would obtain similar overall scores. Groups can then be used to compare and contrast search behaviours and interactions of crowdsourced workers against the ones obtained by correspondent groups of laboratory-based participants. This approach provides a means for comparing search behaviours and interactions between the two user populations.

However, crowdsourcing tools do not usually allow requesters to ask personal questions to users, such as their age, sex, etc. Moreover, it is yet unclear how to judge the truthfulness of answers related to self-perception questions in crowdsourcing environments,[5] such as workers' confidence with search engines and search tasks, their expertise, etc (Feild et al. 2009). Previous studies have observed that workers have strong tendency to not answer truthfully, but rather provide answers they believe the requester is most satisfied with. Platforms such as uTest[6] provide strategies to track the worker demographic in advance of knowing what work is available: this feature is however not available across several crowdsourcing platforms. In AMT, qualification tests may be employed as a source of demographic characteristics. Qualification tests have to be chosen carefully in order to

(1) Not violate the crowdsourcing platform's policies;
(2) Avoid or limit doubts on the trustworthiness of the acquired data;
(3) Yet obtain information that characterises users and their abilities.

To address these points, we propose to use qualification tests based on aptitude or Intelligence Quotient (IQ) tests developed in Psychometrics (Carter 2007). An example of such tests adapted from the Psychometrics literature can be retrieved at http://df.arcs.org.au/quickshare/def16679a0ad6655/Aptitude%20Test.pdf. A further use of this kind of tests is to assess whether workers are suitable to class of information seeking tasks used in the experiments (e.g. domain specific applications). The intuition is that these tests provide a measure of reasoning skills, language knowledge, and problem solving skills of crowdsourced workers, as well as a measure of their attention when performing crowdsourced tasks. Whilst this approach may provide an indication of whether workers provide random answers to our questions or not, it does not provide a strong indication of the trustworthiness of the acquired data. This issue indeed requires further research.

It is yet to be determined whether high IQ scores correspond to higher abilities in solving IIR tasks: this has to be further investigated. However, we expect that there is not a predominant score (or range of scores) among the ones obtained by crowdsourced workers. Conversely, we expect that if the same tests were performed by participants of laboratory studies recruited among the student population of universities, the scores would be predominantly grouped within a high score range, mainly because of the level of education of the participants.

[5] Although similar considerations may apply also to laboratory-based user studies.

[6] http://www.utest.com/, allows requesters to have access to a large population for testing software applications.

2.2 Define information seeking tasks

Information seeking tasks assigned to crowdsourced workers have to be clear and well defined, as no interaction is possible between workers and requesters. Workers are unlikely to perform the cognitive effort required by simulated situations and information seeking tasks, as workers' main goal is to complete tasks as efficiently and rapidly as possible. We suggest that in crowdsourced IIR environments, researchers should explicitly provide the topic that the search will be about, together with a number of specific informational questions the workers are expected to answer.

For example, one of the topics contained in the experiments we report in Sect. 4 is "Australian wines". Once the topic has been assigned, workers are given the opportunity use a search engine we provide for helping them gather information that can assist them answer the following questions related to Australian wines:

- What winery produces Yellowtail?
- Where does Australia rank in exports of wine?
- Name some of Australia's female winemakers.

We argue that posing questions about a specific topic initiates in the workers the search requirements needed by the settings of IIR experiments, eliminating the requirement of simulated tasks used instead in traditional laboratory-based user studies (e.g. see Leelanupab et al. 2009). Our claim is motivated by the fact that in the experimental methodology based on crowdsourcing, the scenario in which the information seeking task is performed is clear: workers get paid for answering a number of questions. To do so, they can find information that assists them answer these questions by searching through the provided IIR system. Topics and questions should be carefully chosen so that answers are not likely to be known, and search needs are thus effectively initiated. Furthermore, we believe the findings observed under these settings may be generalised to scenarios other than "get paid for answering questions by searching" (where for example the monetary reward is substituted by the information gain the user of an IR system experiences).

2.3 Capture interactions

Once topics and questions are assigned, workers can search with the provided IIR system in order to find useful information to formulate answers. It is imperative for the IIR system to capture the interactions between the workers and the system itself (e.g. issued queries, clicked results, time spent in reading/searching, etc). Crowdsourcing platforms, such as AMT, do not provide native tools for capturing these kind of user interactions.

To overcome this issue, two different solutions might be developed:

(I) A web link located within the crowdsourced HIT is displayed to workers, who are redirected by clicking on the link to an external web service that is controlled by the experimenters, and thus able to record workers' interactions. The architecture of a system implementing this schema is sketched in Fig. 1*(Left)*. Crowdsourced workers access the HIT description provided by the crowdsourcing platform (1), e.g. AMT, which provides to the user a link to an external resources (2) that when followed presents the worker with the search interface and service hosted on an external server, controlled by the experimenters. Thereafter the worker interacts online with the external server's services (3 and 4). This solution is inconvenient because it creates payment issues: HITs have to be paid through the crowdsourcing platforms, but

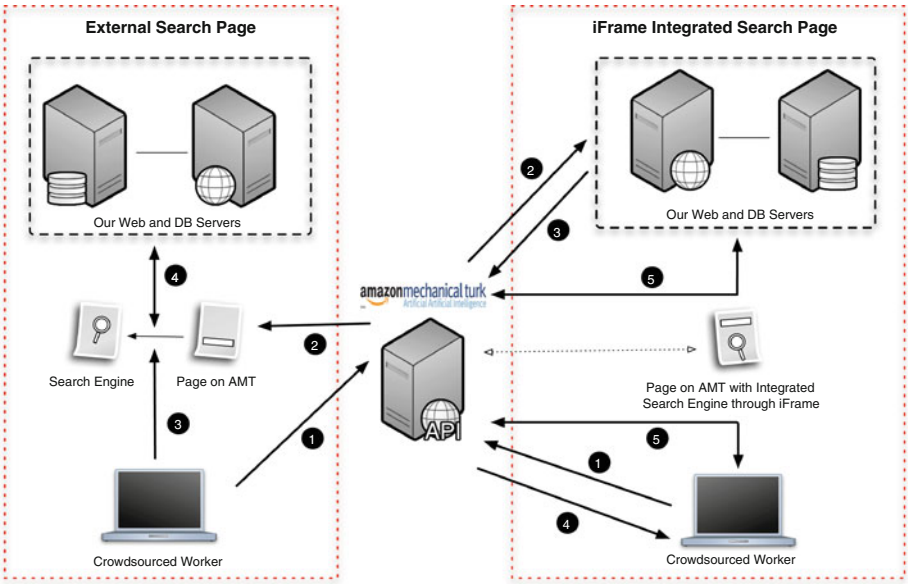


Fig. 1 System architecture based on an external search page that is reached by the worker through a web link made available in the AMT HIT *(Left)* and an iFrame integrated within the AMT HIT interface *(Right)*

because work has been carried on outside the platform itself, the work performed by each worker cannot be automatically identified. The identification has to rely on the worker reporting their worker identification within the external web page, and by assigning a token to each performed action which must then be submitted for the HIT. A similar expedient has been used for example in McCreadie et al. (2011).

(II) Workers are shown the interface of the IIR system through an iFrame which is self-contained within the crowdsourcing platform's HIT interface. Through the iFrame (as shown in Fig. 1*(Right)*), the services offered by the researchers' web server can be provided to the workers, who, when interacting leave an associated identifiable "fingerprint" in the logs of the web services. In this way, interactions can be recorded to be available for further analysis, without explicitly directing workers to external resources. As for solution (I), crowdsourced workers access the HIT description provided by the crowdsourcing platform (1), which however does not present a web link but displays the experimental search interface within the HIT. The search displayed within the HIT is sourced from the experimenters' web services (2 and 3) and is made available to the worker to interact with (4). Subsequent interactions occur then between the worker and the experimenter's web services through the crowdsourcing platform's service (5). This solution effectively overcomes the problems of solution (I).

2.4 Acquire post-search information

Self-perception information about the search task workers just performed can be acquired by means of a questionnaire within a unit of work. Questions can be related to the difficulty

of the task, the level of satisfaction with both system and answers provided, etc. However, little can be said about the truthfulness of the acquired data (Feild et al. 2009). Nevertheless, this problematic issue can be partially addressed by well known techniques, e.g. different phrasing of subsequent questions, so that answers cannot be inferred by the context.

In a crowdsourcing environment, feedback obtained during the post-search questionnaire has the side-effect of being useful in filtering poor worker quality. It would be expected that if the worker has made reasonable effort in completing the task, then answers given in the questionnaire should correlate with features extracted from the search engine's interaction log. For example, if a user provides an answer (whether right or wrong) for a given question, and states that they found the answer using the search interface provided, then there should of course be interaction related to finding that answer. Similarly, if a user did not know the answer before starting the task, and state they found the search task easy yet fail to provide any reasonable attempt answer and perform very little interaction then validity of the work could be questioned.

3 Comparison between crowdsourcing-based and laboratory-based IIR evaluation

In the following, we outline some of the key aspects that differentiate crowdsourcing-based from laboratory-based IIR experiments.

3.1 Heterogeneity

The user population that can be reached through crowdsourcing is often more heterogeneous than that usually available for traditional IIR studies with respect to location, nationality, education, employment, age, gender, language, etc.[7] With respect to this issue, we refer the interested reader to the work of Ross et al. (2010), who performed a demographical study of the workers of AMT, and that of Ipeirotis (2010a, b). It might be argued that the crowdsourced workers who take part in the HITs we describe in this work are self selected through the public crowdsourcing marketplace, meaning that background, environment and skills of the workers' might bias the selection of the HITs workers accept to perform.

3.2 Cost

Crowdsourced IIR experiments are likely to be cheaper than laboratory based ones. For similar experiments we ran during preliminary studies and which involved 58 HITs, we paid an average hourly rate of $1.38 (Zuccon et al. 2011b). In the experiments we report in Sect. 4, where we asked workers to complete 480 HITs uniformly divided in five different pay groups, the average hourly rate was $2.02. As a comparison, the national minimum wage in UK is about $9.35.

[7] It can be argued that the average user of crowdsourcing platforms is reasonably well educated to know English and how computers and crowdsourcing platform work; furthermore, they would have sufficient economic means and geographic access to be using a computer and the Internet.

3.3 Scale

Because researchers can access a large number of workers through crowdsourcing tools, and because of the associated low costs, crowdsourcing often provides the opportunity to reach a higher number of participants for IIR experiments than laboratory-based approaches.

3.4 Users' information quality

While it is often assumed that participants in laboratory-based user experiments provide researchers with correct and detailed information about themselves,[8], the same cannot be assumed for crowdsourced workers. In fact, usage regulations of web-based crowdsourcing platforms often forbid researchers asking personal details of users (e.g. see AMT policies[9].) Furthermore, malicious users may participate in crowdsourced tasks. Finally, crowdsourced workers likely optimise their working strategy for completing tasks, so as to achieve task completion with the minimum effort or within a minimum time.

3.5 Typology of IIR tasks

Traditional IIR experimental paradigms prescribe the creation of simulated work task situations, with participant often required to read instruction sheets that outline the system usage, a simulated situation and an information need the user has to address. This procedure is unlikely to be suitable for crowdsourced workers: previous studies have observed that instructions provided to workers have to be kept short and simple, and workers are unlikely to perform the cognitive effort required by simulated situations. In our method, we suggest assigning a search goal to workers, but do not cast them into domain-specific scenarios.

3.6 Quality of interactions/reliability of interactions

Previous studies suggested that crowdsourced workers tend to complete tasks as efficiently as possible (Feild et al. 2009). Furthermore, malicious workers might submit tasks without actually performing the requested operations. These aspects pose doubts on the quality and reliability of interactions captured through crowdsourcing. Interactions obtained via crowdsourcing should be validated and then compared against those acquired with traditional approaches.

4 Evaluating IIR systems: a case study

In the following we study two different IIR systems instantiated in both laboratory and crowdsourcing settings. Note that the two settings are characterised by different procedures: in fact, due to the constraints of crowdsourcing (e.g. lack of details about the population or sessions for training and interviewing, etc.), it is not always possible to replicate the same procedures in both settings. In the laboratory-based user experiments we decided to adopt standard laboratory-based methods (Kelly 2009) because we do not want to simply simulate the crowdsourcing methods within the laboratory. Instead, we aim to

[8] Researchers select a group of qualified subjects and ask their personal information.

[9] https://requester.mturk.com/mturk/help?helpPage=policies.

compare the two settings, acknowledging the existence of differences in the methodology that go beyond just the location (i.e. laboratory or crowdsourcing platform) of the experiments.

Experimental settings and research questions are described in this section. Results obtained from the two different experiment methodologies are reported in Sect. 5, while they are analysed, compared and contrasted in Sect. 6.

4.1 Research questions

Within the scope of this article, we focus on the following research questions:

RQ1: Can crowdsourcing marketplaces and platforms be used for IIR evaluation? How does the outcome of the crowdsourced evaluation compare with that of the laboratory-based user studies?
RQ2: Are there similar characteristics in terms of search behaviours, interactions, strategies, etc. between laboratory users and crowdsourced ones?
RQ3: Do different payment levels associated with crowdsourced work influence the evaluation results and the evaluation reliability?

4.2 Experimental settings

In both laboratory and crowdsourcing settings, the experiments consist to assign to a user an information seeking task composed on questions on a topic that need to be answered. To help users answer the questions, we provide them with one of our two experimental IIR systems.[10] Users can search during the allocated time and thus gather relevant information that can help them answer our questions.

In our experiments, we considered four tasks, each of them consisting of a different topic. A topic composes a HIT in our experiments, and a single HIT was performed several times by different users in both laboratory and crowdsourcing settings. For each topic, we required users to answer to three questions. Topics and related questions are reported in Table 1, and have been extracted from the TREC 2006 and 2007 Question-Answering tracks (Dang et al. 2006, 2007). The first topic relates to the Pakistani government overthrown that happened in 1999, when former prime minister Nawaz Sharif, accused of plane hijacking and attempted murder, was deposed in a coup on Oct. 12, 1999, by Pervez Musharraf. The overthrown was formally disapproved by a number of countries, including Egypt, Kuwait, Saudi Arabia, Iran. The second topic used in the experiments relates to the 1999 Baseball All-Star Game that was originally scheduled to take place in Milwaukee, Wisconsin, but that effectively took place on July 13 in Boston, Massachusetts, due to delays in the construction of Milwaukee's new baseball arena. The third topic concerns with the US Air Force's B-17 bomber, also know as Flying Fortress, which served during World War II in both the European and Pacific war zones, mainly bombing German and Japanese objectives. The most famous B-17 had been named Memphis Belle, that flew 25 combat missions, plus a last return mission to US. Finally, the last topic we used in the experiments is about Australian wines. In particular we focused on the Yellowtail wine, produced by the Casella Wines Pty Ltd winery, based in Yenda, New South Wales, and on specialist Australian female winemakers. We also asked users to give us information about the export of Australian wine. This of course is susceptible to annual variation; however

[10] The two systems employed in our experiments are described in Sect. 4.4

Table 1 Four search tasks and their questions used in laboratory-based and crowdsoucing-based experiments

T1 (146)	Topic: Pakistani government overthrown in 1999	
	Questions	(T1.1) Who was the nominal leader after the overthrow?
		(T1.2) For what crime was the deposed leader found guilty?
		(T1.3) Which countries formally disapproved of the overthrow?
T2 (161)	Topic: 1999 Baseball all-star game	
	Questions	(T2.1) In what city was the 1999 Game originally scheduled?
		(T2.2) List the official sponsors of the game
		(T2.3) What was the date of the 1999 All-Star Game?
T3 (276)	Topic: B-17 bomber	
	Questions	(T3.1) What was the nickname given to the B-17 Bomber?
		(T3.2) How many missions did the Memphis Belle fly?
		(T3.3) The B-17 bomber was used against which countries?
T4 (279)	Topic: Australian wine	
	Questions	(T4.1) What winery produces Yellowtail?
		(T4.2) Where does Australia rank in exports of wine?
		(T4.3) Name some of Australia's female winemakers

Australia is consistently in the top 10 wine exporters worldwide (7th largest wine exporter in 1999, 6th in 2009, 4th in 2010). These topics were chosen from a larger selection of topics belonging to the Question Answering tracks. This is because in previous experiments they were shown to trigger a larger number of interactions with the IIR systems than other topics (Zuccon et al. 2011a, b). Moreover, the selected topics vary in content and the answers to our questions are unlikely to be well known by all the users. Finally, some topics might present interesting temporal issues for further analysis, while others might present location-biased issues. For example, the answer for question T4.2—"Where does Australia rank in exports of wine?"—is affected by temporal fluctuations associated with the yearly amount of wine Australia exports: while in the 1999 Australia ranked 7th in the wine export market, in 2010 Australia has been the fourth largest wine exporter worldwide. Location-based issues might be associated to topics as T1—"Pakistani government overthrown in 1999"—or T2—"1999 Baseball All-Star Game"—for which users coming from particular countries, i.e. Pakistan and US, might be advantaged in solving the search tasks. An analysis of if and how temporal and location issues affect laboratory and crowdsourcing based IIR evaluation is out of the scope of this article; therefore we leave these for future work.

It is worth noting that topics/questions are characterised by unique factual answers or complex open answers and are indeed of different levels of difficulty and complexity. This opens up two issues:

1. How to assess the correctness of answers given by user; and
2. How to assess the difficulty of topics.

The correctness of answers provided by users and workers can be assessed by contrasting these against the set containing the correct and complete answers. The most reliable but time consuming strategy for assessing the correctness of answers consists of manually annotate each answer as being correct or not. Such manual labelling can be

carried out by one experimenter, to enforce consistency on the assessments; alternatively, each answer can be judged by multiple assessors, and then use a voting strategy based on inter-judges' agreement to establish the final assessment of each answer. By doing so, automatic methods can also be employed. For example, one might resort to (i) use keyword matching, (ii) measure the distance between the language model of the answers with that of the golden set (using for example the Kullback Leibler divergence), (iii) use metrics adapted from the tasks of document summarisation, such as Rouge-N (Lin 2004). Automatic approaches are inevitably affected by errors, while being less time-consuming than manual labelling. Finally, researchers could adapt the crowdsourcing paradigm to the task of assessing the correctness of answers: workers can be provided with the questions and the answers, and are requested to say whether the answer is correct or not, possibly using the same search interface used in the experiments, if needed, so as to give a proof of evidence for their answer. To enforce quality in the collected judgements, majority voting or other quality control strategies[11] could be used by the researchers. As for the automatic assessment, also this solution requires less effort in terms of researcher's time; moreover, although yet affected by noise and errors, it might turn out to be more reliable in terms of assessment precision than automatic assessment strategies.

In the experiments reported in this paper, we performed a manual labelling of all the answers provided by experiments' participants. We further categorise answers as being correct and complete (CC), correct (C), wrong (W) and not given (N). Examples of CC answers are Casella Wines Pty Ltd winery (for question 4.1) and Louisa Rose, Jane Hunter, Pam Dunsford, etc (for question 4.3); while, if question 4.3 was answered naming only one of the female winemakers, then we would judge the answer as being correct (C), but not complete. For some questions, there is no distinction between C and CC: question 4.1 admits only Casella Wines Pty Ltd as answer and is labelled as CC, while the C label is ignored. Furthermore, we considered C and CC answers as being right answers, while W and N answers as being wrong answers.

To gather indicative estimations of the difficulties of the search topics, we have asked our users to rate how difficult it was to answer the questions and indeed to find relevant documents that could have suggested an answer. Ratings were given on a five point Likert scale, where 1 corresponded to the task being very difficult to solve, and 5 being very easy to achieve. From the experiments we conducted, we found that indeed the four tasks had different levels of perceived difficulty. Task difficulty assessments are reported in Table 2, for both laboratory and crowdsourcing experiments (the latter divided at different payment level and aggregated with respect to all rewards). The feedback from the participants suggests that task T3 is the easiest task among those used in the experiments: this is consistent in both laboratory and crowdsourcing based settings, irrespective of the payment levels. Conversely task T2 is felt to be difficult, although crowdsourced workers felt on average this was the most difficult task among all, while laboratory-based users thought task T1 was more difficult.

In our experiments, budget and number of HITs were set with respect to the laboratory-based user experiments, so as to replicate typical laboratory-based IIR studies. Experiments were ran under a constraint budget, as often is the case in laboratory-based IIR evaluation. The settings of the crowdsourcing-based user experiments thus depend on the budget and number of HITs allocated for the laboratory-based user study.

[11] See for example the work of Ipeirotis et al. which presents an algorithm for identifying bias and errors in labelling tasks by assigning a score to each worker so as to represent the quality of their work (Ipeirotis et al. 2010).

Table 2 Averages of the ratings regarding the *perceived difficulty* of tasks given by laboratory participants and crowdsourced workers

Task	Lab.	Crowdsourcing						All
		$0.1	$0.2	$0.3	$0.4	$0.5	All crowd.	
T1	3.04	3.50	2.96	2.88	3.17	2.88	3.08	3.07
T2	3.17	3.00	3.04	2.88	2.58	2.75	2.85	2.90
T3	4.04	3.83	3.83	3.63	3.58	3.63	3.70	3.76
T4	3.42	2.96	2.50	3.13	3.04	3.08	2.94	3.02

Ratings have been collected in the post-search questionnaire using a five-point Likert scale (Difficult–Easy, 1–5). The label "All crowd." refers to the averages obtained considering all the level of payments, while label "All" indicates average values obtained considering both laboratory-based and crowdsourcing-based judgments

To compare and contrast IIR systems in our evaluation framework, we observed the following metrics:

- Temporal length of a session, i.e. amount of time spent completing the task
- Length of a session in terms of number of queries posed to the system in that session
- Number of right and wrong answers
- Number of documents clicked
- Number of documents marked relevant
- Number of result pages examined

We also designed a qualification test to be taken by both laboratory-based participants and crowdsourced workers. The goal of the test was to quantitatively characterise the users of the systems. To this aim, we followed the idea outlined in Sect. 2.1, and we selected 20 questions of which 15 were chosen from a IQ/aptitude test of aptitude test taken from (Carter 2007), while the remaining 5 were extracted from an English TOEFL questionnaire.[12] The complete qualification test is reported at http://df.arcs.org.au/quickshare/def16679a0ad6655/Aptitude%20Test.pdf.

Finally, ethical approval was obtained from the Ethic Committee of the College of Science and Engineering of the University of Glasgow for both laboratory-based and crowdsourcing-based user studies. The form submitted to the institution for seeking ethical approval can be retrieved at http://df.arcs.org.au/quickshare/dbd16ffbe09a2084/Lab_Ethic_Committee.pdf and http://df.arcs.org.au/quickshare/ffb400b5f9807e72/Crowdsourcing_Ethic_Committee.pdf.

4.2.1 Laboratory-based experimental settings

In the laboratory-based user experiments, we adopted a Graeco-Latin Square[13] design for rotating and counterbalancing systems and search tasks (independent variables). According to this experimental design the order of systems and tasks undertaken by the participants are rotated so as to reduce learning effects, which can influence the outcome of the study (dependant variables).

[12] http://www.ets.org/toefl/.

[13] A Graeco-Latin Square is formed by merging two orthogonal Latin square of an $n \times m$ arrangement over two sets of variables, e.g. systems and tasks.

Table 3 The experimental design of our laboratory-based IIR evaluation follows a Graeco-Latin square rotation for systems (S1–S2) and tasks (T1–T4), involving 24 users (U1–U24)

User		Systems and tasks rotation				
		Slot 1	Slot 2		Slot 3	Slot 4
U1	10 min Training	S1, T1	S2, T2	5 min Break	S1, T3	S2, T4
U2		S2, T2	S1, T3		S2, T4	S1, T1
U3		S1, T3	S2, T4		S1, T1	S2, T2
U4		S2, T4	S1, T1		S2, T2	S1, T3
…		…	…		…	…
U24		S2, T4	S1, T1		S2, T2	S1, T3

According to Shadish et al. (2001), the nature of human behaviour is one of the problems affecting a user study. This, in particular, results from natural learning aptitude, by which humans can learn how to handle a system and solve a task. Thereby human behaviour in one condition will influence their behaviour in another. In other words, results of subsequent experiments most likely will be better than the results of earlier experiments.

Table 3 shows an example representation of the Graeco-Latin Square design, where we randomly assigned participants to different rows. By doing so, all users (U1, …, U24) will perform search on all evaluating systems and in all given search tasks, but in different orders of unique randomised pairs of system and task. Note that the Graeco-Latin Square design cannot eliminate the learning as, for example, task T2 mostly follows task T1, but this design can equally distribute its impact of system order and partially of task order across all experimental conditions (Kelly 2009; Leelanupab 2012). Although a complete randomisation of the order of tasks can be achieved, it is somewhat impractical to follow in laboratory settings[14].

We allocated a budget of £240 (about $400) for performing the laboratory-based evaluation of our two IIR systems. We wanted to test each system on 4 different tasks, and we decided to constrain the average time necessary for performing all the required procedures associated to each task to maximum 20 min. We decided to pay each participant £10 (equivalent to about $16.60) for taking part in the study, which extended for maximum 80 min. The payment is in line with the UK minimum wage, which is the common payment rewarded to user study participants by the University of Glasgow. Therefore, given our budget, we set our number of participants to 24.

We invited 24 University of Glasgow undergraduate and postgraduate students (16 males and 8 females), aged 17–39 (mean= 26.32 years old) to take part in our laboratory-based IIR study. Their academic backgrounds vary from psychology and philosophy to statistics, and biomedical and computer science. Half of participants were English native speakers while the rest were advanced and intermediate English speakers.

We conducted five sessions a day and completed the experiment within five working days. Participants started the study by performing an aptitude test, which followed the qualification test taken by crowdsourced users. They had 10 min to perform this test. Afterwards, participants were asked to answer an entry questionnaire regarding their personal background and search experiences. The answers show that participants are familiar with online search services (e.g. Google, Bing, Wikipedia, etc.), and consider

[14] This is due to several constraints in laboratory-based user experiments such as the limited number of experimenters and participants as well as time and budget.

these systems easy and simple to use, often finding what they search for. Thereafter we gave the users 10 min to get familiar with our search interface, and perform a training task.

For each participant, we allocated 13 min per task for the following steps (1) perform the searches required to answer our questions, (2) write the actual answers in the associated form, and (3) provide quantitative feedbacks by answering a post search questionnaire upon the completion of each task. The post search questionnaire consisted of a set of 12 semantic differentials questions on Five points Likert scales, and concerned with participant's impressions about the performed search experience and the used system. A final open-question was inviting participants to leave comments. A 5 min break was given to participants at the end of the second slot, to comply with ethical regulations.

An exit questionnaire was also provided at the end of study (i.e. after all 4 tasks were completed), where participants were asked to compare the two different systems, and indicate which one they preferred, providing a motivation for their choice. A final short interview session was allocated to enquire about participants' search strategies when performing the tasks.

4.2.2 Crowdsourcing-based experimental settings

The crowdsourcing-based evaluation was carried out on the Amazon Mechanical Turk (AMT) in early 2011, and was set as follows. We created batches containing the same 96 pairs of systems/tasks as those used in the laboratory-based user study. Each pair formed a HIT in the crowdsourcing platform. Workers could select to perform up to 4 HITs: our search service was in charge to assign to participants a task they did not yet perform, and to rotate the systems accordingly, i.e. a worker taking 4 HITs will do so by using two times system S1, and two times system S2. This expedient did not guarantee that a users takes all 4 HITs, however it guarantees that (i) users take one single task only one time, and (ii) users taking an even number of HITs use both systems for the same amount of time.

We required that users participating to our experiments had an AMT's acceptance rate of at least 95 %, that means, maximum the 5 % of the work that have previously completed on the crowdsourcing platform had been rejected. Initially, we planned to require a compulsory qualification test based on an aptitude test for users to be eligible to perform our HITs. However, from preliminary experiments we carried out in previous works (Zuccon et al. 2011a, b), we noticed that the presence of such requirement had the effect of slowing down the completion of the batches.[15] Apparently, in fact, workers were not keen to undertake an unpaid survey that lasted 10 min, such as the qualification test, without knowing if they could have subsequently performed the HITs, or if they would have even enjoyed to perform them. Since the aim underlying the introduction of our qualification test was not for filtering out workers that achieved low scores, but instead aimed to provide us a quantitative characterisation of the crowdsourced participants, we made the qualification test not compulsory for taking our HITs. However, we informed the workers that could have taken the qualification test, promising a bonus of $0.1 if they were doing so. We waited 3 days from the first HIT a user performed before checking if he also performed the qualification test or not. If the qualification test was not performed, we planned to remove the HIT and the related interactions from our logs: however, we found that participants taking our HITs were also always performing the qualification test afterwards, probably attracted from the bonus we promised.

[15] The observation of qualification tests slowing down batch-completion time is consistent with the findings reported by Alonso and Baeza-Yates (Alonso and Baeza-Yates 2011).

The structure of the HITs is described in Sect. 4.3 However, note that post search questionnaires were issued to workers at the bottom of the HIT interface (under the iFrame containing the search interface). We could not require workers to fill entry and exit questionnaires due to the restrictions imposed by the workflow of the crowdsourcing platform. We instead added an extra question to the post search questionnaire asking workers to rate the system's performance: we use this indication to compare system preferences against those obtained in the laboratory settings. Note that an alternative experiment architecture may have been used instead to collect answers to these questionnaires. Qualification tests may have been used as entry questionnaire, an exit survey that resembles the exit questionnaire may have been associated with the award of a bonus.

In previous work, Mason and Watts showed that in crowdsourcing settings there are little or no relations between the payment awarded to workers per HIT and the quality of the performed work (Mason and Watts 2009). Similar evidences have been reported by Potthast et al. (2010), who showed that payment had effect on completion time but not on result quality, in the context of crowdsourced plagiarism detection in documents. However, Kazai reports opposite findings: payment does matter, and an higher pay is often guarantee of better work (Kazai 2011). To study whether payment has an impact on the quality of the work, and in particular on the system evaluation indications that are obtained from the crowdsourced interactions, we decided to create several numbers of batches where we varied the reward amount of each single HIT. Each batch of HITs was thus characterised by a different reward amount. We also wanted to maintain a 2:1 ratio between the budget spent in the laboratory-based IIR evaluation and that to be spent in the crowdsourcing based evaluation, therefore allocating about $200 for the crowdsourcing experiments. Following our constraints and needs, we created one batch of HITs for each of the following amounts: $ 0.1, $ 0.2, $ 0.3, $ 0.4, $ 0.5. This amount to a total spend of $156, including the fees charged by the crowdsourcing platform (10 % of the HIT reward). This left us with $44 to be spent in awarding bonuses to workers, such as those promised for taking the qualification test. Batches with different payments were ran at different times, but workers were allowed to select HITs belonging to different batches, provided that they were not performing the same task twice: this condition was enforced by our system. Further bonuses were promised (and awarded) if all 4 tasks were completed by a worker ($0.05 bonus) and if the quality of the work was satisfactory (a one off bonus of $0.1). We judged that a set of HITs were performed in a satisfactory manner if the worker spent more than 450 s working on the HIT, issuing more than 2 queries if answers were not known, and provided at least one answer.

We had in total 304 different workers participating to our crowdsourced system evaluation. With 304 unique users working on our HITs we spent $33.44 in bonuses related to the qualification tests, while $12.38 were spent in awarding the remaining typologies of bonuses. The total costs of the crowdsourcing experiments amounted thus to $201.82, which was slightly higher then the budget we initially planned to spend. Note that not all the HITs were accepted. However, we tended to accept the highest number of HITs as possibly, as the development and assessments of methodologies dedicated to filter out low quality workers is out of the scope of this article. Specifically, we rejected only those HITs that were clearly fabricated, i.e. where the answers provided did not match with the correspondent check-boxes indicating whether an answer was found or not, and its source.

In Fig. 2 we provide the breakdown for reward level of the distribution of our 304 workers per country from which they accessed the crowdsourcing platform: this information was captured in the logs by our system. The clearly show that payment levels can be used as a tool to select the country from which workers access the HITs: high rewards

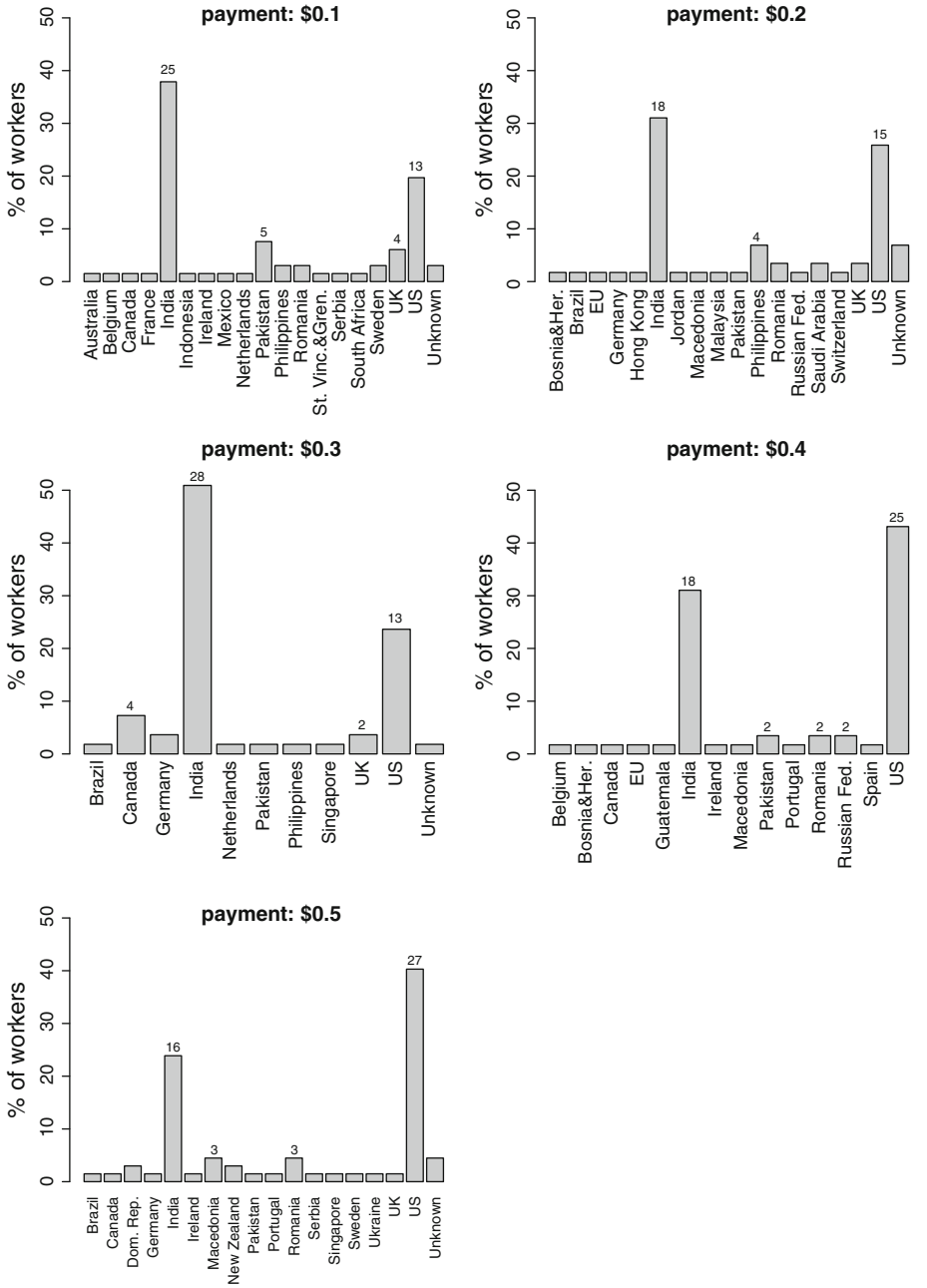


Fig. 2 Distribution of the crowdsourced workers for country of access

(i.e. $\geq$\$0.4) attract a significant larger amount of US-based workers than lower rewards. While, HITs characterised by a low reward (i.e. \$0.1–\$0.3) are more often performed by workers connecting from India, Pakistan and Philippines.

Note that differently from laboratory-based user experiments, in the crowdsourcing-based user study we could not provide a training session for the workers, where to get familiar with the search interface. We do not believe this caused major problems, because the interface was similar to those used by popular search services. We decided however to provide a training session in the laboratory-based user experiments to conform to the standard protocol of laboratory-based user experiments. In fact, our goal is to compare the traditional protocol with that devised for crowdsourcing: if training sessions were removed from the laboratory-based user experiments, we would have not fairly compared the two protocols.

Furthermore, we could not control workers' fatigue and in particular award them a 5 min break between each pair of HITs, as instead it has been done in the laboratory-based user study.

4.3 HIT description

While in laboratory settings users are presented with two different windows (each positioned in a separate screen) containing the system interface and the data forms, respectively, in the crowdsourcing settings all this information has to be enclosed within the interface of an AMT HIT. The AMT crowdsourcing platform makes available a number of HIT formats. For the purpose of our experiment we use the simple form builder features of an AMT-hosted HIT that contains our search engine and collects worker answers and feedback (shown in Fig. 3). The form data collected from workers during the HIT is used to generate a comma-separated value text file which is later merged with the external query-log data.

The HIT contains the same information given in the laboratory settings to the participants, but it is designed to be as intuitive as possible. A bullet-point summary (A) of the task guidelines, cautions and bonus criteria is given first. Workers are explicitly told to find their answers using the embedded search engine. Failure to do this can be easily detected by the lack of query-log interactions.

The task questions (B) are introduced by setting a context ('Australian Wine' in the case shown in figure) and then providing the three specific questions that the worker should attempt to answer. Query log analysis in earlier work (Zuccon et al. 2011a, b) indicated that crowdsourced workers tend to copy and paste the question text contained in the HIT into the search box, rather than take time to naturally formulate a query to address the described information need. To avoid this undesired behaviour we presented questions as image-based text, therefore preventing text selection.

The search engine is embedded in the HIT using an iFrame window (C). JavaScript is used to pass the worker and assignment identifiers, along with the task being answered as to allow query-log data to be related with completed HITs. Because the AMT platform has no built-in method of excluding individual workers from a HIT, the search engine imposes the constraint of disallowing the same worker from performing the same task more than once.

The text box labelled as (D) allows a user to input their answer and identify the source for each question. Finally, a post-search questionnaire (E) is displayed to gather feedback. Upon submission, client-side JavaScript validation ensure that all necessary form fields in (D) and (E) have been input or selected. If there is missing input, the user is given a pop-up summary of the additional information needed for submission.

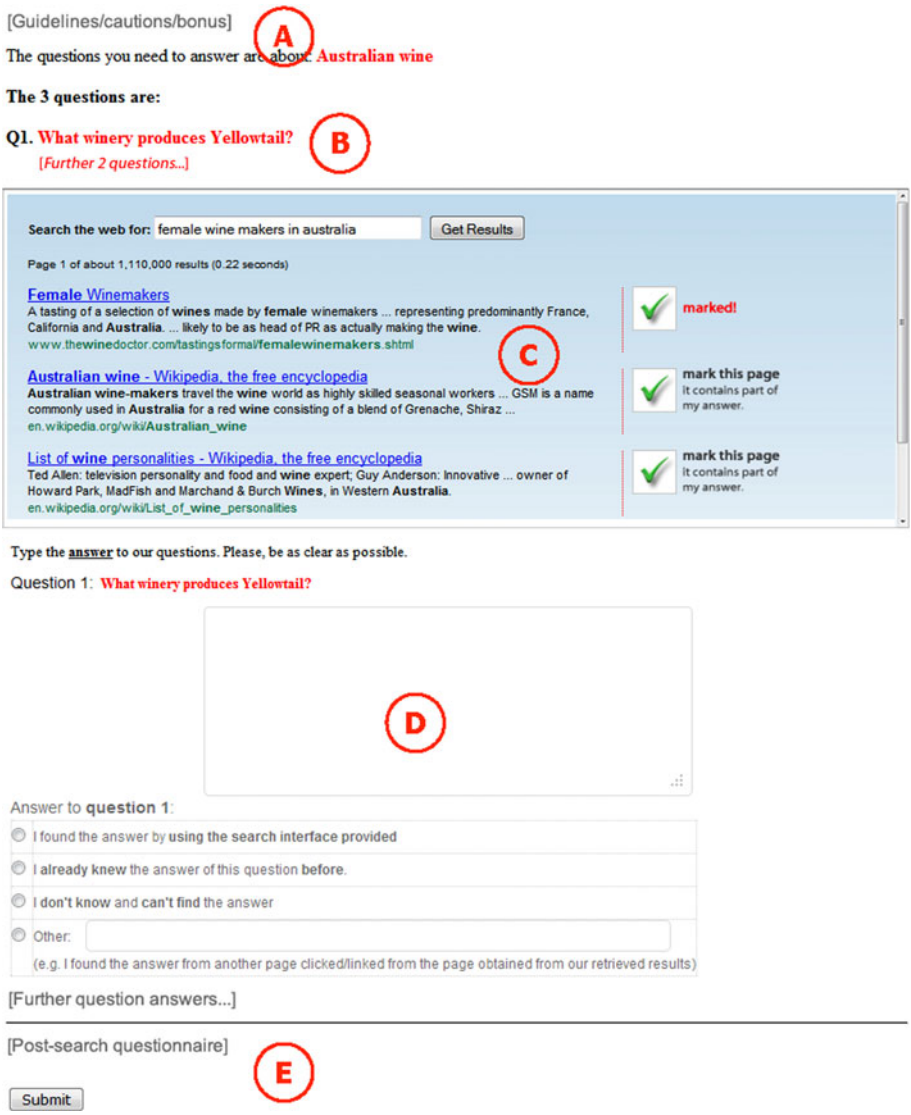


Fig. 3 A sketch of the HIT layout we used in our crowdsourcing experiments

4.4 Experimental systems

In the following we describe the common architecture that is used to develop the two IIR systems. The systems are also based on a common interface, which is detailed in Sect. 4.4.2; while the retrieval algorithms underlying the systems are presented in Sect. 4.4.3.

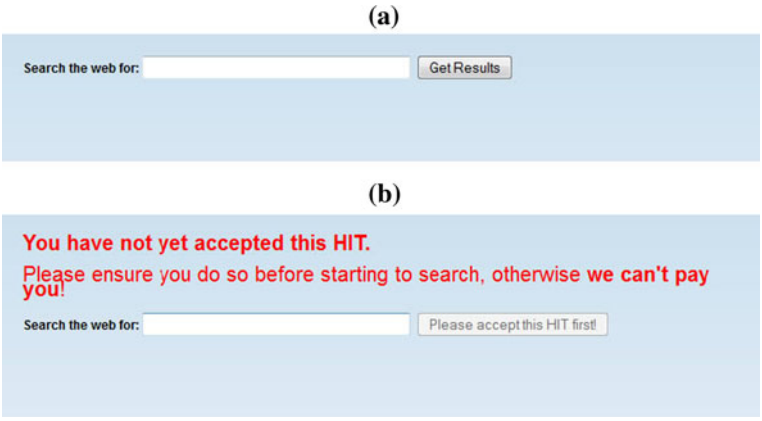


Fig. 4 The search interface: initial interface with empty search box. Image **a** refers to the interface presented to laboratory-based participants, while image, **b** refers to the interface displayed to crowdsourced worked in case they had not yet accepted the HIT

4.4.1 Common system architecture

Both systems rely on the web-service provided by the Microsoft Bing API[16] for web results. When queries are submitted to a system, they are routed to the Bing Web Search service, which returns a set of results that are thereafter elaborated by each system, based on the process applied by the specific retrieval algorithm employed. For each returned result, we display the same information that is commonly displayed by Bing, i.e. document title, snippet, url. To comply with Bing restrictions and regulations, and to guarantee adequate time-response to our systems, we fetched only a maximum of 50 results in answer to each of the submitted queries.

Issued queries, dwelling time, documents clicked, explicit relevance feedbacks (i.e. button "Mark this Page" in the search interface), and other interactions are recorded on a database, which is also used to store additional information collected through questionnaires and qualification tests.

4.4.2 Common system interface

The systems tested in our experiments use a common interface, shown in Figs. 4 and 5. The interface imitates those of popular search engines, such as Google and Bing. It consists of an initial search box, where users can type their first query (Fig. 4a). In the crowdsourcing settings, this interface is slightly modified by inserting a message reminding the workers they cannot interact with the system until they have accepted to perform the HIT (Fig. 4b). This expedient is necessary because otherwise crowdsourced workers would be allowed by the AMT service to interact with the HIT before accepting the HIT itself. This compromises the possibility to associate interactions performed by workers to particular search sessions recorded in the logs.

Clicking of the "Get Results" button takes the users to the list of search results, which is composed of documents' titles, snippets and links (Fig. 5). The colour schema and the layout of the displayed results follow those of popular search engines. In each result page, a

[16] http://www.bing.com/developers.

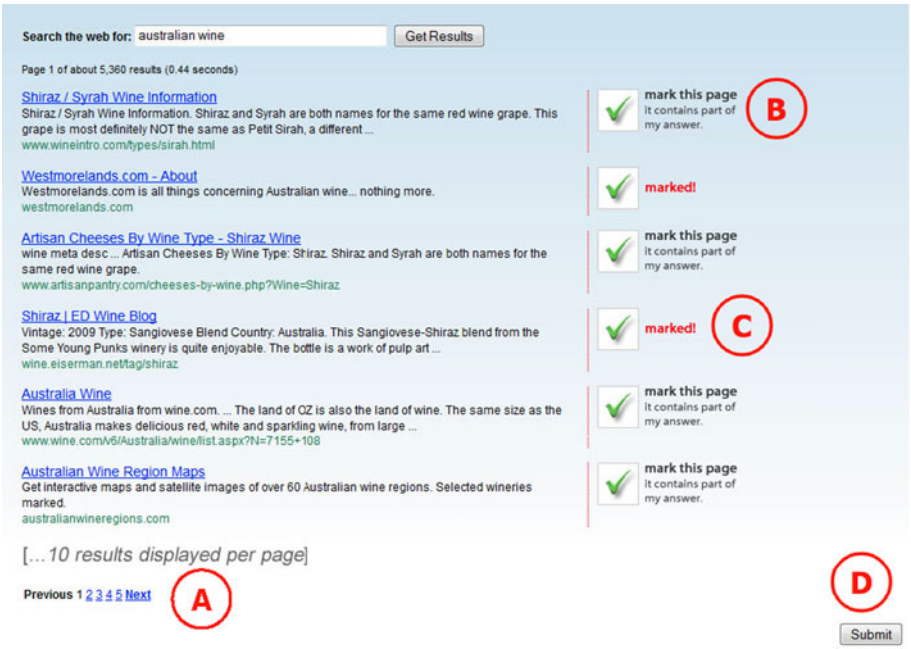


Fig. 5 The search interface: the page containing the results for the submitted query

maximum of 10 results are displayed. Users can navigate through the result pages using the numbered links at the bottom of the interface (A): however, a maximum of 5 pages (and 10 results per page) can be reached for each query.[17]. By clicking on a search result, a pop-up window opens to display the content of the retrieved document. Additionally, on the right hand-side of each result a button (i.e. "Mark this Page" button) gives users the option to mark a document as being helpful to answer the questions (B), i.e. to be relevant to the search task. Once this button has been clicked, its aspect changes into "marked" (C): the action is recorded into the logs and cannot be further revised. Finally, the button on the lower-right hand-side corner of the interface allows users to terminate their search sessions, and submit the HITs (D).

4.4.3 Retrieval strategies

Each of the two tested systems is characterised by a different retrieval strategy. The first system, S1, passes the query submitted by the user to the Bing search services, obtains the list of retrieved documents and returns them to the users. Conversely, the second system, S2, submits the original query to the Bing search engine obtaining the correspondent search results, but it also issues up to 4 additional queries that the Bing search service suggests as being related. Once retrieval results for the related queries are obtained, they are fused together with the document ranking obtained for the original query. The fusion follows a round-robin procedure, where the first result corresponds to the first result obtained from the original query issued, the second result corresponds to the first result retrieved in answer to the first query suggestion provided by Bing, and so on.

[17] Recall that due to Bing API's limitations we only retrieved maximum 50 results per query

System S2 mimics a system that strives to diversify the search results given an initial query. Specifically, we developed the system borrowing the idea of multiple query submission from the work of Santos et al. (2010). In contrast, system S1 represents a traditional Web retrieval system.

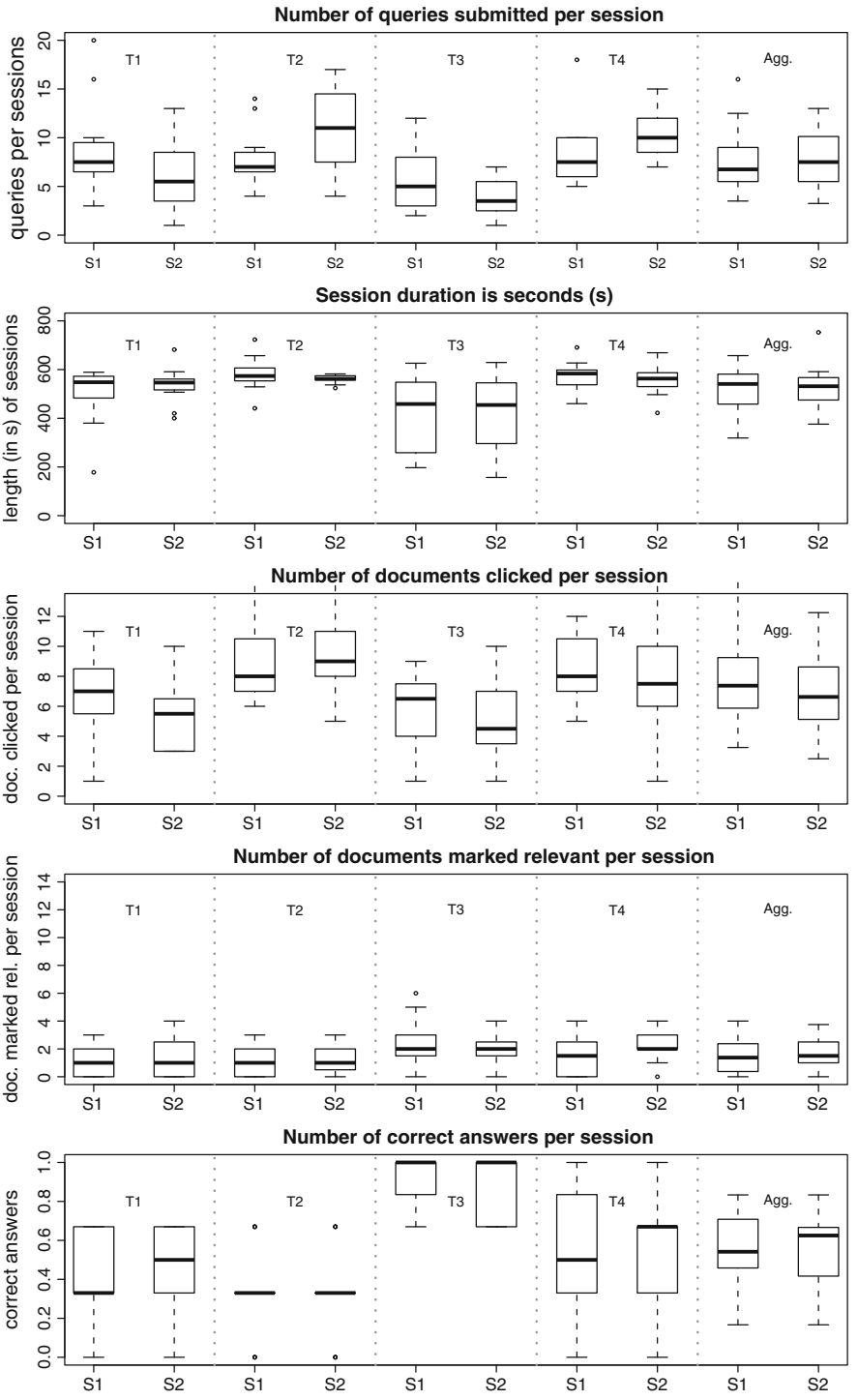
5 Results of the case study

5.1 Laboratory-based results

In Table 4 we report the number of correct answers (from 3 questions) per search task provided by participants of the laboratory experiments on the two IIR systems. The average score (Avg. Score) of how many correct answers the participants can provide is presented in each corresponding task and summary row. Figure 6 also presents these statistics. The

Table 4 Number of correct answers per sessions provided to each of the three questions contained in the four experimental tasks. The final row of the table reports the total number of correct answers over all 4 topics. No statistical significant differences were measured when performing a two-tailed t-test with confidence level $p > 0.95$

Laboratory-based evaluation			
Task	Correctness	S1	S2
1	0/3	1	1
	1/3	7	5
	2/3	4	6
	3/3	0	0
	Avg. score	1.25 (± 0.62)	1.41 (± 0.67)
2	0/3	2	2
	1/3	8	8
	2/3	2	2
	3/3	0	0
	Avg. score	1.00 (± 0.60)	1.00 (± 0.60)
3	0/3	0	0
	1/3	0	0
	2/3	3	4
	3/3	9	8
	Avg. score	2.75 (± 0.45)	2.67 (± 0.49)
4	0/3	1	1
	1/3	5	3
	2/3	3	6
	3/3	3	2
	Avg. score	1.67 (± 0.98)	1.75 (± 0.87)
	0/3	4	4
over	1/3	20	16
All	2/3	12	18
Tasks	3/3	12	10
	Avg. score	1.67 (± 0.95)	1.71 (± 0.90)


Number of queries submitted per session
queries per sessions
Session duration is seconds (s)
length (in s) of sessions
Number of documents clicked per session
doc. clicked per session
Number of documents marked relevant per session
doc. marked rel. per session
Number of correct answers per session
correct answers
T1
T2
T3
T4
Agg.
S1
S2

◀ **Fig. 6** Laboratory-based evaluation: Box-plots for the number of queries submitted per session obtained with respect to the two different search systems we evaluated (S1 and S2), over 4 different tasks (T1, ..., T4). The box-plots labelled with "Agg." refer to the aggregated data, i.e. the average over all the 4 tasks

answer data shows that according to the laboratory evaluation there are no sensible differences between the two systems. System S2 helps user to provide more correct answers, in particular for task T1 and T4: however differences are not statistically significant. In particular participants provided on average more correct answers for tasks T1 and T4 when using system S2 rather than system S1, although the variance associated to this measurements is very large. The increase in the correctness rate is achieved by examining less documents (i.e. compare the number of clicked documents per session for tasks T1 and T4 as shown in Table 5), but seeing fewer documents does not affect the number of documents that are marked relevant between the two systems.

Similarly, in Table 5 we report the amount of interactions we recorded in the laboratory-based evaluation, divided for each individual search task (T1, …, T4) and averaged over all tasks (Avg.). The trends regarding the number of issued queries on the topics T1 and T4 do not follow the pattern that characterised the previous statistics. Specifically,

Table 5 Values of the collected measures during the laboratory-based evaluation of two IIR systems (S1 and S2), over 4 search topics (tasks $1, \ldots, 4$), and averages for each system over all 4 topics. In the column marked s.s. we reported whether statistical significant differences were measured: the presence of the symbol $\surd$ signals the presence of statistical significant differences. These were measured using a paired t test, with confidence level $p > 0.95$

Laboratory-based evaluation				
Measure	Task	S1	S2	s.s.
Number of queries	1	8.75 (±4.90)	6.00 (±3.36)	$\surd$
	2	7.92 (±2.94)	10.75 (±4.29)	$\surd$
	3	5.75 (±3.39)	3.92 (±1.88	$\surd$
	4	8.33 (±3.60)	10.42 (±2.50)	$\surd$
	Avg.	7.69 (±3.84)	7.77 (±4.23)	$\surd$
Temporal session length	1	502.25 (±118.67)	536.08 (±73.26)	
	2	580.42 (±68.57)	599.17 (±136.95)	
	3	412.83 (±151.31)	426.83(±144.44)	$\surd$
	4	572.17 (±59.07)	559.17 (±63.48)	$\surd$
	Avg.	516.92 (±123.30)	530.31 (±125.05)	
	1	7.08 (±2.75)	5.42 (±2.39)	$\surd$
Document clicked	2	9.00 (±3.25)	9.33(±2.74)	
Per session	3	6.25 (±3.41)	5.08 (±2.71)	$\surd$
	4	8.75 (±3.19)	8.00 (±3.54)	
	Avg.	7.77 (±3.27)	6.96 (±3.31)	$\surd$
	1	1.08 (±1.00)	1.50 (±1.51)	$\surd$
Document marked	2	1.17(±1.03)	1.33 (±1.07)	
Relevant	3	2.42 (±1.68)	1.92 (±1.16)	$\surd$
Per session	4	1.50 (±1.38)	2.25 (±1.06)	$\surd$
	Avg.	1.54 (±1.37)	1.75 (±1.23)	$\surd$

users issue on average more queries when using system S2 than when using S1 in task T4. While, for task T1 the average number of issued queries for system S2 is lower than that for S1. The number of issued queries is not consistent also with respect to the other two tasks (i.e. T2 and T3), where less queries issued with one system does not translate into lower or higher (average) number of correct answers provided. Instead, users are able to answer to the questions of tasks T2 and T3 with equal precision when using systems S1 or S2. Moreover, there appears to be a relationship between the number of submitted queries and the number of clicked documents. In fact, if more queries are submitted using one particular system, then more documents are examined using that same system, and vice-versa. This may suggest that search results are consistently explored regardless of the overall number of queries issued. The observed relation is true on average and for tasks T1, …, T3, but not for task T4, although in this case the number of clicked documents presents a high variance when participants used system S2.

Table 6 (column "Lab") summarise the judgements made by participants about the effectiveness of the two systems in helping them solve the tasks. Overall, when adopting the laboratory-based methodology, no statistical significant differences are evidenced between the two compared systems with respect to the correctness of the answers provided by participants. This may suggest that (1) the two systems were not found statistically different when helping users uncover correct answers, or (2) the population sample was too small to achieve statistically significant results with respect to this indicator. However, system S2 appears to help more the users when trying to solve tasks T1 and T4. Statistically significant different search behaviours, i.e. number of issues queries, relevant documents marked, etc, can be identified between search interactions that involve the two different systems. Furthermore, by examining the users' ratings regarding system effectiveness in supporting their information need tasks it can be noted that system S2 is generally preferred to system S1 (Table 6, column labelled "Lab"): differences are significant also in this case.

5.2 Crowdsourcing-based results

In Fig. 7 we report the number of submitted queries per session with respect to the five batches of different payments (i.e. \$0.1, \$0.2, \$0.3, \$0.4, and \$0.5) in the crowdsourcing experiments. Each plot illustrates the statistics with respect to each search system obtained on different search tasks (i.e. T1,...,T4) and those obtained by aggregating the data of all the tasks. On average, there seem to be no apparent differences between the two systems when considering how many queries users issued: only when the payment is set to \$0.5, users queried less when using system S2 than when using S1. However, when single tasks

Table 6 Average values of the system effectiveness assessments with respect to the two different IIR systems studied in our experiments. The label "s.s." refers to the statistical significance analysis: specifically, $\surd$ indicates statistical significant differences as measured by a two-tailed t-test with level of confidence $p > 0.95$

System	Lab.	Crowdsourcing						All
		\$0.1	\$0.2	\$0.3	\$0.4	\$0.5	All crowd.	
S1	3.46	3.42	3.23	3.17	3.38	3.06	3.25	3.32
S2	4.00	3.31	3.19	3.38	3.50	3.38	3.35	3.41
s.s.				$\surd$	$\surd$	$\surd$	$\surd$	$\surd$

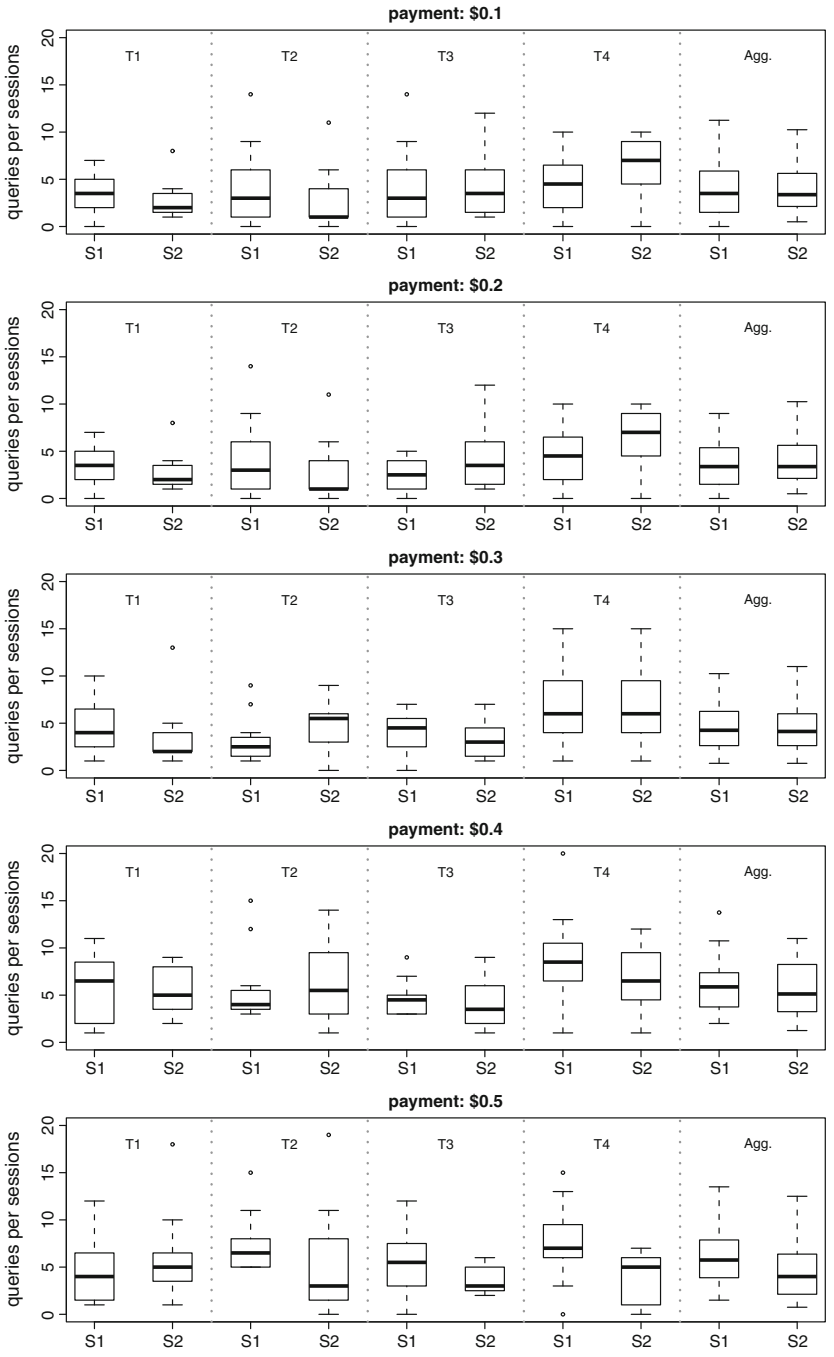


Fig. 7 Crowdsourcing based evaluation: Box-plots for the *number of queries submitted* per session obtained w.r.t the two different search systems we evaluated (S1 and S2), over 4 different tasks (T1, …, T4) at different payment levels ($0.1, …, $0.5). The box-plots labelled with "Agg." refer to the aggregated data, i.e. the average over all the four tasks

are considered for each level of payment, the data suggests that the number of submitted queries corresponds, to some extent, to the assessments of perceived tasks difficulty provided by those same workers (Table 2).

For payments $\leq$\$0.2, the trends exhibited by the number of submitted queries statistics are similar on all four tasks and highly correlated to the task difficulty from user's perception. Specifically, workers rated tasks T1 and T2 as being moderately difficult (ranked 2nd and 3rd), and the number of queries issued for these two topics have similar patterns with respect to both mean and variance. On the contrary, workers that performed higher paid HITs (i.e. \$0.3, \$0.4, and \$0.5) rated tasks T1 and T2 as being more difficult but exhibit different query-issuing behaviours (in terms of number of issued queries) from those exhibited by worker of the lower rewarded HITs. This might be the case because different payments attracted different populations of workers, as Fig. 2 shows. These workers might have different background knowledge on the search topics, or different search abilities in terms of query formulation and text understanding. lengths

The lengths of the search session performed in the crowdsourcing experiments are summarised in Fig. 8. No distinct pattern is evident among payments with respect to the duration of information seeking tasks. Results unveil that some workers do not spend any time using the search system (some sessions lasted zero seconds or slightly more). This might be due to workers already knowing the answers to our questions, or not willing to genuinely perform the task. We found that workers that were paid more than \$0.1 per HIT spent on average less time searching with system S2 than with S1. This might suggest that system S2 allows workers to faster find a equal or higher number of correct answers than what they could get when using S1 (see Fig. 11). Although we found the opposite behaviour when analysing the batches of HITs rewarded with \$0.1, we also found that session lengths in this batch presented a very high variance.

Figures 9 and 10 report the number of documents respectively clicked and marked relevant within a session. The first statistics shows that over all the four tasks an equal or slightly higher number of documents are examined when using S1 than when using S2. This does often correspond to an increase in the number of relevant documents that are marked per session. This consideration does not, however, apply in the case of the interactions obtained when paying \$0.1 per HIT, where a slightly higher number of relevant document are found when using S2.

It is interesting to note how the previous observations relate with the number of correct answers provided by workers when using different systems and at different pay regimes. This data is reported in Fig. 11. When considering the \$0.1 batches, more relevant documents were marked when using S2, rather than S1. However in these settings, the number of correct answers provided by the workers who used S2 is lower than their fellow workers that used S1. When payment increases, the number of correct answers obtained using S2 increases as well. Moreover, S2 is found to help workers provide an equal or larger number of correct answers than S1 in the pay levels $\geq$\$0.3.

The previous findings are in agreement with the system effectiveness assessments provided by the crowdsourced workers, which are reported in Table 6. When low payment levels are considered, i.e. $\leq$\$0.2, users felt that system S1 was more effective than S2 in assisting them during the tasks, although no statistical significant differences between the two systems are found. While, when payments increases, system S2 is felt to be better than S1, and statistical significance is found in the differences of assessments. The differences in assessments and search behaviours that we found at different levels of payment might be due to the differences that are found in the worker population when different rewards are paid (see Fig. 2).

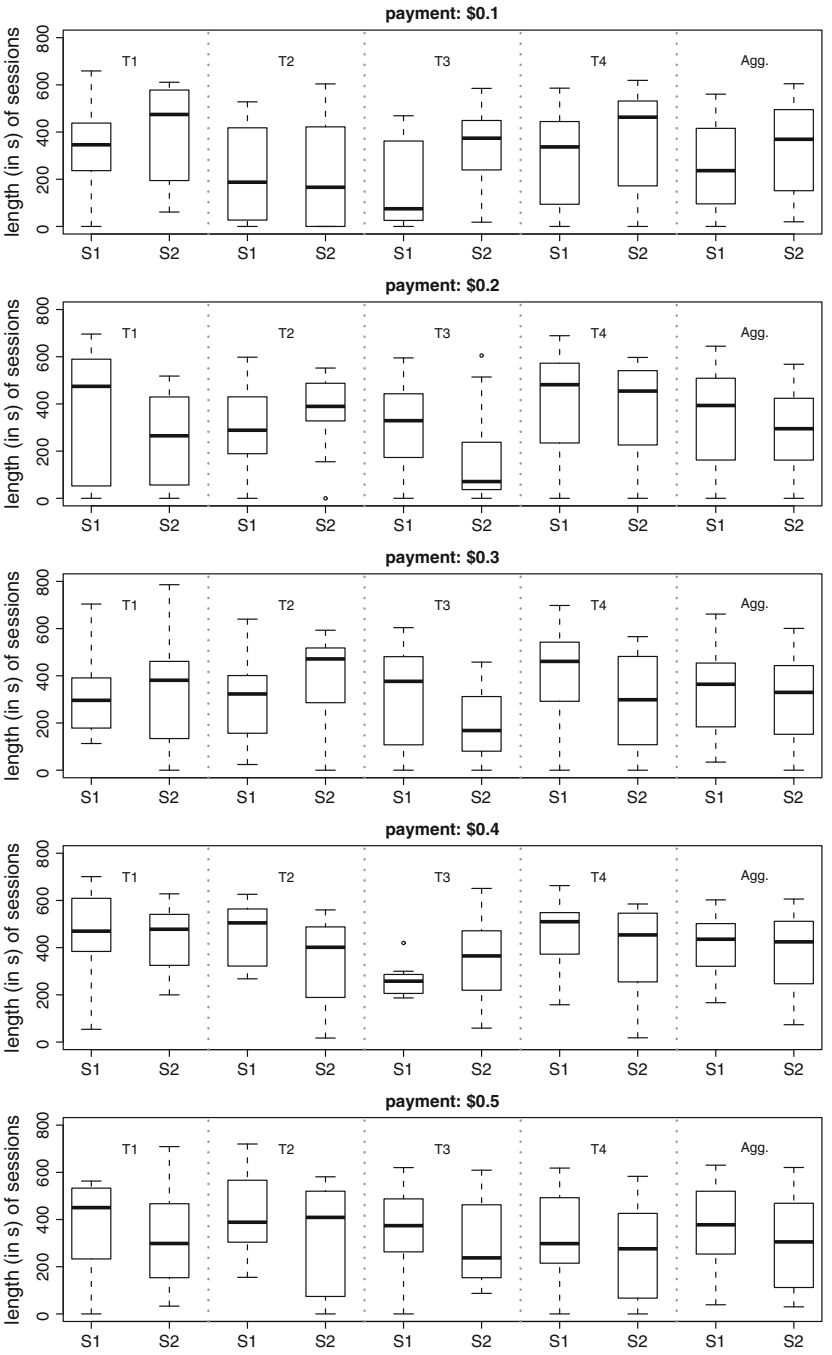


Fig. 8 Crowdsourcing based evaluation: Box-plots for the *session duration is seconds (s)* obtained with the w.r.t different search systems we evaluated (S1 and S2), over 4 different tasks (T1, …, T4) at different payment levels ($0.1,…, $0.5). The box-plots labelled with "Agg." refer to the aggregated data, i.e. the average over all the four tasks

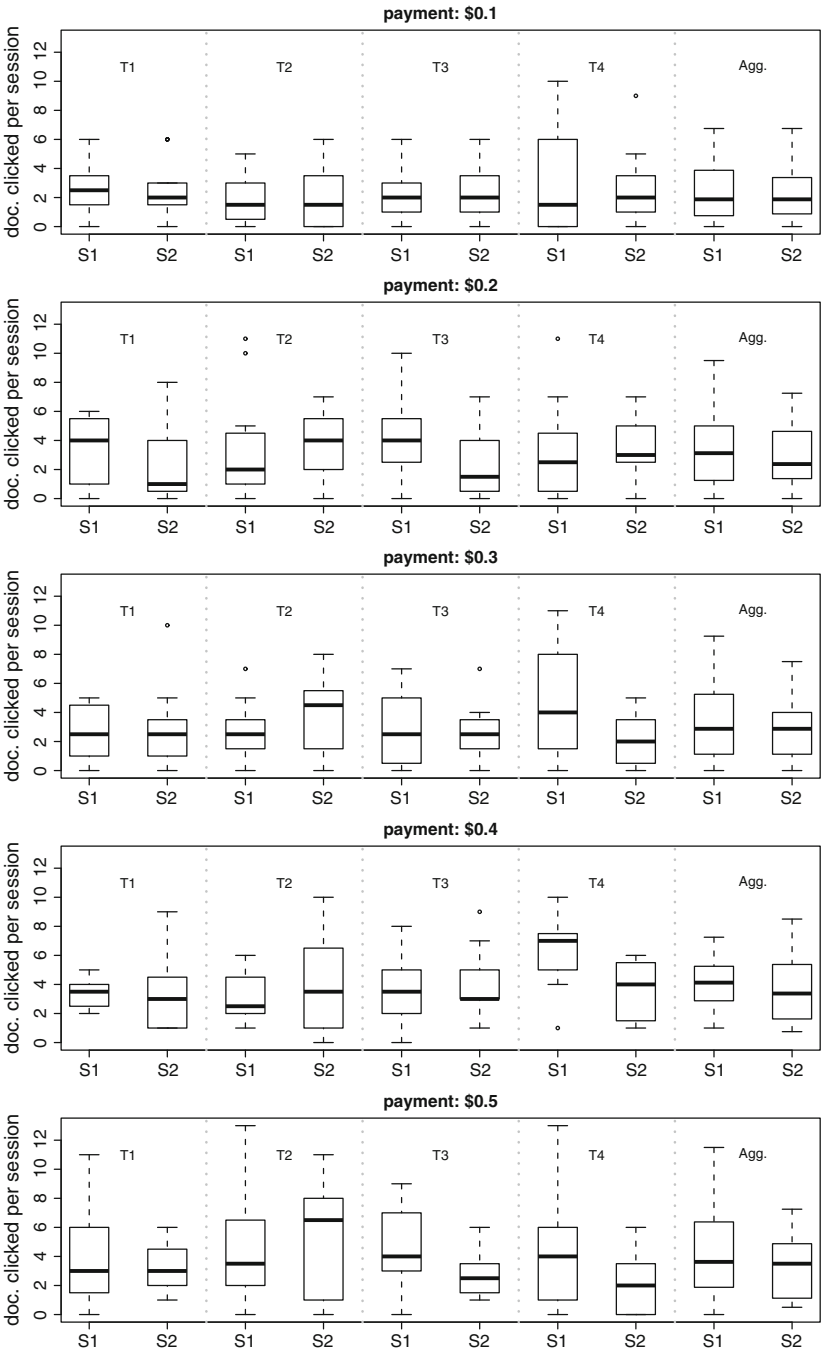


Fig. 9 Crowdsourcing based evaluation: Box-plots for the *number of documents clicked* per session obtained w.r.t the two different search systems we evaluated (S1 and S2), over 4 different tasks (T1,…, T4) at different payment levels ($0.1,…, $0.5). The box-plots labelled with "Agg." refer to the aggregated data, i.e. the average over all the four tasks

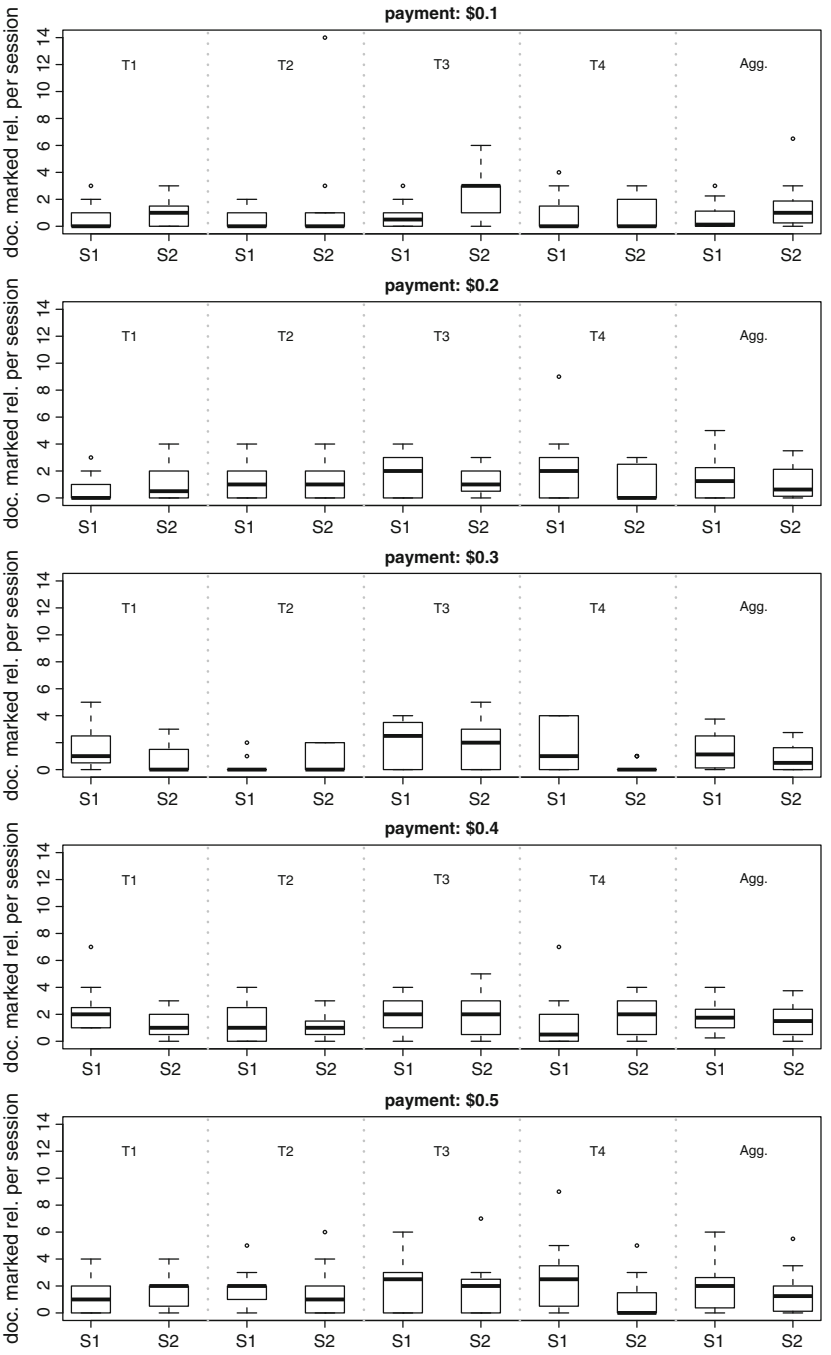


Fig. 10 Crowdsourcing based evaluation: Box-plots for the *number of documents marked relevant* per session obtained w.r.t the two different search systems we evaluated (S1 and S2), over 4 different tasks (T1,…, T4) at different payment levels ($0.1, …, $0.5). The box-plots labelled with "Agg." refer to the aggregated data, i.e. the average over all the four tasks

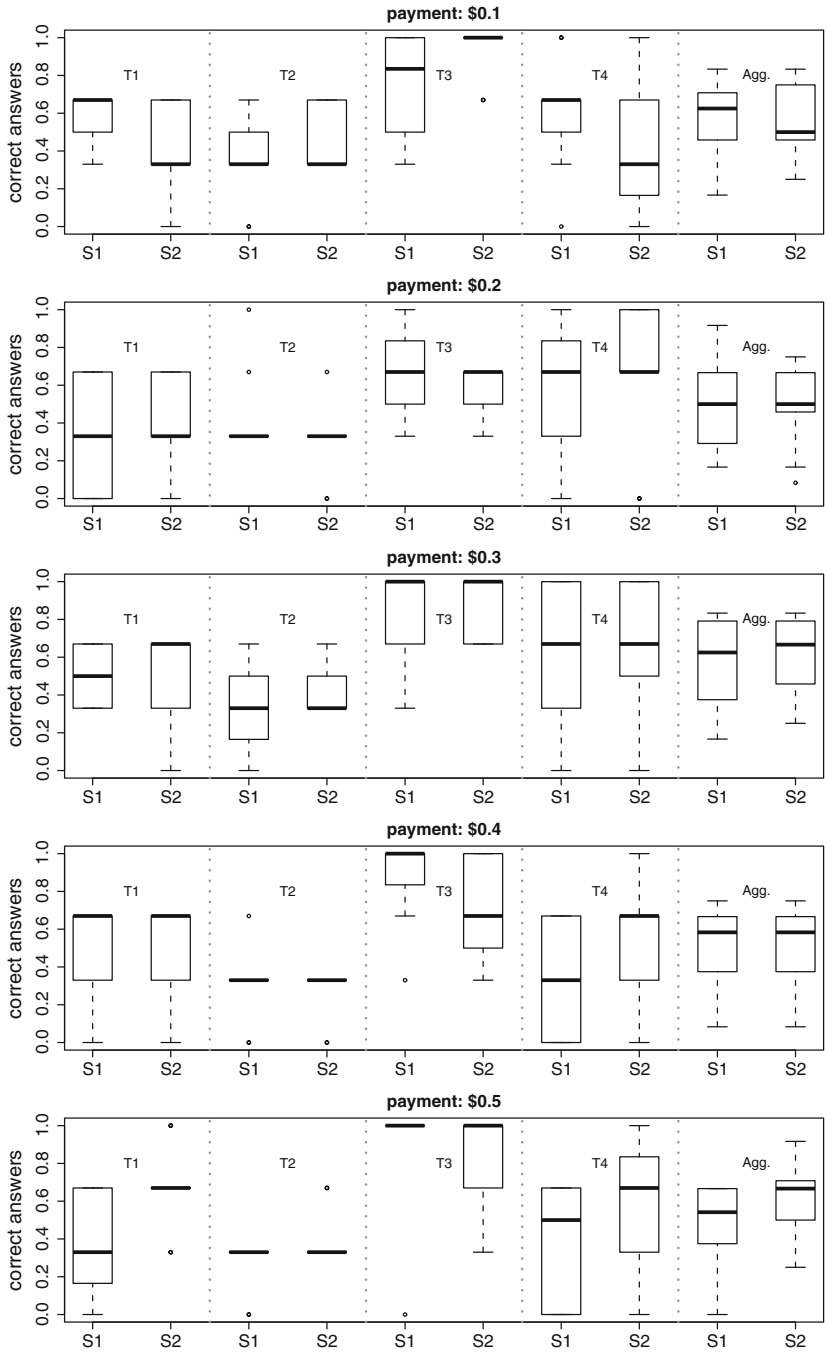


Fig. 11 Crowdsourcing based evaluation: Box-plots for the *number of correct answers* per session obtained w.r.t the two different search systems we evaluated (S1 and S2), over 4 different tasks (T1,…, T4) at different payment levels ($0.1,…, $0.5). The box-plots labelled with "Agg." refer to the aggregated data, i.e. the average over all the four tasks

From the results of the crowdsourcing-based evaluation, we can conclude that when the payment is set to be equal or higher than \$0.3 per HIT the captured interactions and assessments suggest that system S2 does a better work than S1 in supporting the workers in the tasks, and it is generally preferred to S1. Most of the evidences that support this claim are based on statistical significant differences among the interactions generated using either of the systems. When lower payments are considered (i.e. $\leq$\$0.2), these findings are often subverted, but no statistical significant differences are found between the statistics associated to the two systems. Two interpretations can be drawn from this finding. Firstly, as we shown in Fig. 2, HITs with payments $\geq$\$0.3 attracted a different population than HITs with payments $<$\$0.3 (North America and Western Europe versus Asia). This may suggest that mainly workers based in North America and Western Europe prefer system S2 over S1; while workers based mainly in Asia prefer S1 over S2. Secondly, we have observed that HITs with rewards $<$\$0.3 are characterised by a lower number of correct answers than those collected for HITs with payments $\geq$\$0.3. It may then be posited that preferences collected for HITs with payments $\geq$\$0.3 are more likely to be a trustworthy indication of system effectiveness.

However, note that if the statistics collected for all the payment levels are combined, system S2 results once again to be recognised as more effective than S1 (Table 6), as well as it supports workers in finding more correct answers than its counterpart system.

6 Comparison of laboratory and crowdsourcing findings

6.1 Comparing qualification scores

In Fig. 12 we report the distribution of the qualification scores obtained by participants to our studies in laboratory settings (Fig. 12a) and in crowdsourcing settings (Fig. 12b). Participants that took part in our study in laboratory settings prevalently obtained a score of 15 out of 20 (more than 30 % of participants). Scores higher than 15 are rare within laboratory participants: only the 8.3 % of the user population obtained scores higher than 15, and no participant obtained scores higher than 17. However, the qualification score of 9 out of 20 appears to be a lower bound for laboratory participants, as only less than the 8.5 % of that user population obtained lower scores.

The findings are different when crowdsourcing settings are considered. While the majority of the users also obtained a score of 15 out of 20 in the tests ran through the crowdsourcing platform, more than the 28 % of the crowdsourced user population

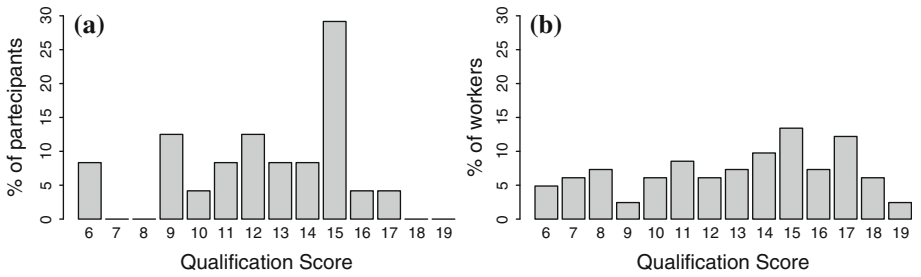


Fig. 12 Qualification scores distribution for the laboratory-based participants (**a**) and crowdsourced workers (**b**)

achieved scores higher than 15, with the 8.5 % of workers achieving the top scores of 18 and 19 out of 20. Similarly, in the crowdsourcing settings there appear to be more users that obtained scores lower than 9 than in the laboratory settings. In fact, scores lower than 9 are obtained by about the 18.3 % of the total worker population of the crowdsourced study.

These findings suggest that the crowdsourced workers are a heterogeneous user population based on the scores obtained from the qualification tests, which aimed at assessing users' background and skills. Conversely, it appears that laboratory-based participants can be categorised into two main score groups, that of scores of about 15 and that of score of about 9.

6.2 Comparing answer correctness

The last plot of Fig. 6 reports the amount of correct answers given to our questions by participants of the laboratory-based user study. Similarly, the plots in Fig. 11 represent the amount of correct answers provided by crowdsourced workers, divided by payment levels. In the laboratory-based user study we found that users gave a slightly higher number of correct answers when they used system S2. The same trend is found when workers are paid \$0.5 per HIT (and differences are statistically significant), although when they are paid between \$0.2 and \$0.4 the number of correct answer does not seem to depend much upon the system that is used. While, when users are paid just \$0.1, system S1 seems to support better their needs, as they provide more correct answers when using S1 (no statistical significant differences are found though). However, while on average workers paid \$0.1 and \$0.2 give the same amount of correct answers provided by laboratory participants (considering both systems), we found that higher paid users gave a higher number of correct answers than the participants to the laboratory study.

6.3 Comparing system effectiveness assessments

Users assessments about the effectiveness of the two IIR systems in supporting their information tasks have been reported in Table 6. We have shown that participants to our laboratory-based user study rated system S2 as being more effective than system S1 in supporting their needs, but no statistical significant differences have been exhibited. This result contrasts with what found in crowdsourcing settings when workers are paid \$0.2 per HIT, or less. Specifically, such workers judged system S1 being more effective than system S2, although no statistical significant differences between the two systems with respect to these judgements has been found. However, if the amount of money paid per HIT is increased, workers tend to judge S2 as being more effective than S1 to assist them in solving their tasks. In these cases, the differences in systems' judgements are statistically significant. This might be due to workers paying more attention to just provide correct answers rather than qualitative feedbacks in tasks with low payments, as this was felt to be the main goal of the HIT. While, higher paid workers might have put equal effort in finding correct answer and providing more resonate feedbacks. This intuition is strengthened by the fact that the number and the quality of feedback received through the optional "leave a comment" text box in the HIT increased with increasing HIT reward. To confirm this interpretation, we contacted a small set of workers[18] (we had in total 19 workers answering our request) belonging to each payment level few days after they completed the HITs them assigned. Specifically, we asked to freely tell us what their strategy in solving the tasks,

[18] Workers has been contacted using the API service made available by AMT.

and whether they gave the same importance to answer questions and fill in questionnaires, or if they felt one was more important than the other. Their answers confirmed our initial intuition.

When the system effectiveness assessments obtained via crowdsourcing with respect to all the payment levels are aggregated, the differences between the judgements obtained for payments of \$0.1 and \$0.2 with respect to those obtained in the laboratory study disappear. That is, when judgements obtained in the crowdsourcing settings are combined, system S2 is rated as being significantly more effective than S1 for supporting users' tasks: this confirms the finding obtained during the laboratory based evaluation.

6.4 Comparing search interactions

When comparing the number of queries issued by laboratory participants and crowdsourced workers, we find that the former submitted on average more queries than the latter (compare the first plot in Fig. 6 and the plots of Fig. 7). This behaviour is consistent irrespective of the level of payment per HIT. However, the trend found in laboratory-based user experiments when the data from all four tasks is aggregated (i.e. on average users submitted the same amount of queries for both systems) is confirmed in crowdsourcing-based experiments. Furthermore, the fewer queries submitted by crowdsourced workers does not correspond to a decrease in the number of correct answers given. Specifically, when payments between \$0.3 and \$0.5 are considered, crowdsourced workers provide more correct answers than users in laboratory settings.

A notable difference between the interactions recorded during our laboratory and crowdsourcing-based user study is the duration of the sessions (see the second plot of Fig. 6 and the plots of Fig. 8). In fact, laboratory-based sessions lasted all about the same amount of time (580 s circa) for the majority of the topics, with very little variance. The only exception to this trend is topic T3, which witnessed some users finishing after just 200 s, for both systems. In contrast, the data collected with respect to session duration in the crowdsourced study shows a higher variance, and shows that the systems were used on average for less time than in the laboratory-based user study. Specifically, in all the batches, regardless of the amount paid per HIT, there were sessions that lasted zero or a few seconds. This may be explained by: (i) sessions of users that knew already the answers and therefore did not perform any search, or (ii) sessions where users did not perform any task and did not provide any answer, but just clicked the check-box associated with "I don't know" and submitted the HIT,[19] or (iii) a mix of the two previous cases. However, the correctness of the answer has not been influenced by the shorter length of the sessions and the presence of zero length sessions.

For reasons similar to those discussed in the previous paragraph, differences are found between the number of documents that are examined (clicked) by laboratory-based participants and crowdsourced workers (compare the third plot of Fig. 6 and the plots of Fig. 9). This suggests that crowdsourced workers tend to examine fewer documents than laboratory-based users, and possibly rely more on the summary of the document that is provided in the snippets. In the crowdsourced studies, the lower amount of clicked documents as well as the lower amount of issued queries might suggest that crowdsourced participants tend to interact less with the result list, and less often reformulate their queries. However, the majority of the crowdsourced workers genuinely attempted to solve their tasks, as witnessed by the amount of correct answers they provide. It is also interesting to

[19] Remember we did not implement filters that exclude this kind of behaviour.

notice that although the workers submitted fewer queries and examined fewer documents, they clicked on a higher number of relevant documents[20] (see the fourth plot of Fig. 6 and the plots in Fig. 10). This might be explained by the fact that crowdsourced workers aim to achieve efficiency with regard to maximising income and might find it more efficient to gather answers or decided to open documents by reading the result snippets to gather. While, in laboratory settings, user may be prone to open a large amount of documents because participant are sure of their pay (regardless of completing the tasks) and may instead feel obliged to interact with the system as long as the allocated session lasts.

7 Summary of findings

In answer to our research questions (Sect. 4.1), we found that:

A1: The use of crowdsourcing platforms for IIR system evaluation is a violable alternative to laboratory based user studies. It was found that evaluations based on the two methodologies lead to similar conclusions about the effectiveness of the employed IIR systems (Sects. 6.2 and 6.3). However, the cost of the crowdsourcing based study was half that of the laboratory based user study, yet collecting five times more data.
A2: Differences in search behaviour and interactions were recorded during the laboratory based user evaluation and the crowdsourcing based evaluation (Sect. 6.4). In particular, it was observed that crowdsourced workers interacted less with the systems, issuing on average less queries, clicking less documents, and in general spending less time using the systems. However, the overall trends observed when comparing the two systems in the laboratory user study are also found when adopting the crowdsourcing methodology (e.g. users submit a similar amount of queries, regardless of whether they use system S1 or S2), and that the fewer interactions recorded did not translate into a decrease in the correctness of the answers.
A3: Payment levels above $0.1 show similar trends with respect to many indicators used in this study (e.g. correct answer, clicked documents, etc), and lead to similar conclusions regarding systems performance, suggesting that reliable information can be obtained when paying more than $0.1. It was found that payments of $0.1 attracted low quality workers, that in general interacted less with the systems (e.g. number of documents clicked (Fig. 9) and number of documents marked relevant (Fig. 10)). Similarly, the effectiveness of the system, as evaluated using data obtained with payment of $0.1, is not consistent with higher payment levels or in the laboratory based user study.

8 Related works

Interactive information retrieval (IIR) focuses on how IR systems meet the information needs of users and how the users interact with *systems* and *information*, including information seeking behaviours and experiences. Traditional IR evaluation abstracts users and employs a simple question (i.e. whether relevant documents are retrieved for a given query) based on the classic evaluation model. Conversely, IIR evaluation is directly related to users and employs multiple methods of data collection and measures to assess system

[20] As judged by themselves.

performance, interactions, and usability. For example, questionnaires and interviews are used to enquire about the search experiences to obtain users' qualitative feedback. Queries, search results, click-through data, and time spent on searching/browsing/assessing documents can be acquired by logs of user interactions. Furthermore, by employing the evaluation methods proposed for IIR, particular features of interactive search systems such as personalisation and query suggestion can be studied individually or as a whole. Users and the context of their search are taken into account in IIR evaluation studies as determinants of retrieval success. We refer the interested reader to the work of Kelly for a complete overview of these aspects (Kelly 2009).

8.1 Laboratory-based IIR evaluation model

Borlund introduced a de-facto framework for the evaluation of IIR systems, where researchers employ users to study elements of systems and cognitive perspectives (Borlund 2003). Her studies of evaluation measures and simulated work tasks using a short cover story have contributed much to the foundation of IIR evaluation.

Typical IIR studies take place in laboratory settings, where researchers are able to control the experimental environment and variables. The impact of one or more experimental variables can be isolated, and the results that are obtained are thought to be reliable due to the control of the experimental variables and the repeatability of the experiments. Although laboratory-based user studies are useful, they are often criticised because they are too artificial, and do not represent real life search scenarios (Kelly 2009). In fact, users' search behaviour may have a higher chance of being contaminated or biased by the experimental design or by the researchers who excessively observe users during the experiments. Besides, experiment data can only be collected for small numbers of users, tasks and systems. The small size of the collected sample and its inherent population bias may limit the generality of the studies' findings. The use of complimentary methods for data collection, such as field studies and ethnographic observations together with log analysis, provides in-depth information of how systems support users in the search process (Grimes et al. 2007).

Finally, Kelly suggested that an additional method for exploratory search system evaluation is to create a *living laboratory* on the Web, which requires thorough development of longitudinal research designs with a larger number of users (Kelly et al. 2009). However, one major problem undermining the success of this latest approach is how to bring researchers and users together. This issue clearly motivates us to use the crowdsourcing marketplace to obtain larger numbers of search interactions and qualitative feedback from a arguably more heterogeneous user population.

8.2 Crowdsourcing for relevance evaluation

Several works have focused on capturing relevance assessments or labels for documents using crowdsourcing. Alonso *et al*. crowdsourced relevance assessments by asking workers to evaluate the relevance of results retrieved by a geographical IR system (Alonso et al. 2008). While, Alonso and Mizzaro compared crowdsourced relevance judgements against the correspondent judgements obtained by TREC assessors (Alonso and Mizzaro 2009), and found that crowdsourced workers provide relevance assessments that are similar to those collected by TREC assessors. They also found that crowdsourced workers were able to detect errors in the relevance assessments given by TREC assessors.

Crowdsourcing has also been used to gather relevance labels for a collection of digital books (Kazai 2011). Similarly, within TREC, crowdsourcing has been employed to gather relevance assessments for the TREC 2010 Blog Track (McCreadie et al. CSDM). Alonso and Baeza-Yates outlined design principles and methodologies for effectively crowdsourcing document relevance judgements (Alonso and Baeza-Yates 2011). In particular, they focused on issues regarding budget for experiments, number of relevance labels to be captured per document, experiment design and schedule, data collection and interface design. The work of Grady and Lease instead uses the human factors involved when relevance assessments are obtained for documents through crowdsourcing (Grady and Lease 2010).

These works are different from our study because:

1. Previous works focused on gathering labels (and in particular, relevance labels) for documents, while we capture a whole interactive search session, consisting of issued query and query-reformulations, clicked documents, depth of result exploration, time spent on documents, etc.
2. In previous works data is gathered and subsequently used to evaluate a system offline; instead, in this study we evaluate IIR systems online, without relying only on relevance assessments, but also on the analysis of search performances and search behaviours.

Other studies have employed crowdsourcing to go beyond labelling and investigate how crowdsourcing can be used to evaluate systems. For example, Arguello *et al*. showed how crowdsourcing can be used to gather preferences about which vertical (i.e. specialised search service) is better to show to users given a pre-defined query (Arguello et al. 2011), and the presentation of different verticals by search systems can be effectively evaluated.

Finally, this study is intimately connected with the proposal made in Zuccon et al. (2011b) and the further investigation in Zuccon et al. (2011a), where it has been suggested to use crowdsourcing to capture user interactions with search engines. In this work, we bring that proposal one step further: crowdsourcing is used to capture search sessions interactions and user feedback, which is then used to evaluate, compare and contrast IIR systems.

9 Directions of future work

Several issues have yet to be investigated: we outline the most important.

First, it is not clear what is the optimal payment for rewarding crowdsourced workers. While a higher payment corresponded to a major decrease of the batch completion time, as observed in Mason and Watts (2009), higher payments reflected only a small increase in answer correctness. Likewise, other statistics appear to be similar among different levels of payment. This consideration does not however apply for the lowest payment level used in our study ($0.1), which often resulted in different trends to those obtained from higher rewarded interactions.

Second, methods have to be devised to be able to filter out HITs of poor quality, e.g. those submitted by bots, for example. In our experiment, we considered valid all the HITs for which there was not a mismatch between the answer provided and the related checkboxes corresponding to whether an answer was answered or not, and where the answer was found. More sophisticated data quality assessments are needed to minimise the chance of obtaining low quality work.

Third, it is very important to consider the objective of crowdsourced workers, that is to achieve efficiency with regard to maximising income. The motivation to complete HITs in as little time as possible means that for more complex work such as the search engine interaction, natural behaviour may be adversely affected. It is therefore important to not only design HITs that can be completed efficiently by workers, but also to consider methods to ensure more natural interaction behaviour. For example, as a result of our experience we implemented image-based questions to stop question copy and pasting, and force cognitive effort in query formulation. Likewise, we introduced submission validation to ensure questionnaires were fully completed, as well as used established questionnaire techniques to avoid workers being able to infer the context of questions without properly reading them.

Fourth, the crowdsourcing approach is characterised by a inherent limitation: researchers cannot reliably gauge some of the personal information of the workers (e.g., gender, age, background, etc.). This can be a big hindrance to performing a conclusive user study of IR and IIR systems.

10 Conclusions

In this paper we employed crowdsourcing beyond its typical use for labelling tasks (Alonso et al. 2008; Kazai 2011; McCreadie et al. 2011). In order to do this, we have proposed and formalised a novel experiment methodology for IIR system evaluation based on crowdsourcing. We have shown how our proposal can be effectively instantiated in the well known crowdsourcing platform of Amazon Mechanical Turk. We have compared and contrasted, both qualitatively and quantitatively, our proposal with the traditional IIR evaluation methodology of laboratory based user studies. An evaluation of two different IIR systems was conducted using both methodologies, and the results were compared. These showed that the crowdsourcing-based IIR experimental methodology provides an evaluation of IIR systems as valuable as that provided by laboratory-based users studies. This was achieved with half the cost of laboratory-based studies, yet collecting five times more data. These findings suggest that the crowdsourcing-based IIR experimental methodology could be used as an alternative to laboratory-based user studies, allowing more cost effective experiments that lead to a larger amount of collected data. However, we believe that a more effective strategy to experimentally evaluate the effectiveness of IIR systems is to combine both methodologies. With this respect, we suggest using laboratory-based user experiments as pilot study for fine tune systems, evaluation tasks and procedures. Thereafter, systems can be tested using the crowdsourcing-based experimental methodology, which allows the access to a larger and more heterogeneous user population and the collection of a larger amount of data regarding systems and search behaviours.

Acknowledgments The authors are thankful to the anonymous reviewers for their suggestions.

References

Alonso, O., & Baeza-Yates, R. (2011). Design and implementation of relevance assessments using crowdsourcing. In: P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee, & V. Mudoch (Eds.), *Advances in information retrieval, volume 6611 of lecture notes in computer science* (pp. 153–164). New York: Springer.

Alonso, O., & Mizzaro, S. (2009). Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *SIGIR '09: workshop on the future of IR evaluation*.

Alonso, O., Rose, D. E., & Stewart, B. (2008). Crowdsourcing for relevance evaluation. *SIGIR Forum, 42*, 9–15.

Arguello, J., Diaz, F., Callan, J., & Carterette, B. (2011). A methodology for evaluating aggregated search results. In: P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee, & V. Mudoch (Eds.) *Advances in information retrieval, volume 6611 of lecture notes in computer science* (pp. 141–152). New York: Springer.

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research, 8*(3), 152. http://www.doaj.org/doaj?func=abstract&id=88950.

Carter P.J. (2007) *IQ and psychometric tests*. London: Kogan Page.

Dang, H. T., Kelly, D., & Lin, J. (2007). Overview of the trec 2007 question answering track. In *Proceedings of the text REtrieval conference*.

Dang, H. T., Lin, J., & Kelly, D. (2006). Overview of the trec 2006 question answering track. In *Proceedings of the text REtrieval conference*.

Feild, H., Jones, R., Miller, R. C., Nayak, R., Churchill, E. F., & Velipasaoglu, E. (2009). Logging the search self-efficacy of amazon mechanical turkers. In *SIGIR 2009 work on crowdsourcing for search eval*.

Grady, C., & Lease, M. (2010). Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with amazon's mechanical turk*, CSLDAMT '10, (pp. 172–179). PA, USA: Stroudsburg. Association for Computational Linguistics.

Grimes, C., Tang, D., & Russell, D. (2007). Query logs alone are not enough. In *Workshop on query log analysis at WWW*.

Ipeirotis, P. G. (2010a). Analyzing the amazon mechanical turk marketplace. *XRDS, 17*, 16–21

Ipeirotis, P. G. (2010b). Demographics of mechanical turk. NYU working paper no. ; CEDER-10-01. Available at http://hdl.handle.net/2451/29585, March 2010.

Ipeirotis, P. G., Provost, F., & Wang, J. (2010). Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, HCOMP '10, (pp. 64–67). New York, NY, USA: ACM.

Kazai, G. (2011). In search of quality in crowdsourcing for search engine evaluation. In P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee, & V. Mudoch (Eds.) *Advances in information retrieval, volume 6611 of lecture notes in computer science* (pp. 165–176). UK: Springer.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3*(1–2), 1–224.

Kelly, D., Dumais, S., & Pedersen, J. (2009). Evaluation challenges and directions for information-seeking support systems. *Computer, 42*(3), 60–66.

Leelanupab, T. (2012). *A Ranking framework and evaluation for diversity-based retrieval*. PhD thesis, University of Glasgow.

Leelanupab, T., Hopfgartner, F., & Jose, J. (2009). User centred evaluation of a recommendation based image browsing system. In *Proceedings of the 4th Indian international conference on artificial intelligence* (pp. 558–573). Citeseer.

Lin, C. Y. (2004). Rouge: a package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization, ACL 2004*. Spain: Barcelona.

Mason, W., & Watts, D. J. (2009). Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD workshop on human computation*, HCOMP '09, (pp. 77–85), New York, NY, USA: ACM.

McCreadie, R., Macdonald, C., & Ounis, I.: Crowdsourcing Blog Track Top News Judgments at TREC. In M. Lease, V. Carvalho, E. Yilmaz (eds) *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the 4th ACM international conference on web search and data mining (WSDM)* (pp. 23–26). Hong Kong, China, February 2011.

Over, P. (1997). Trec-6 interactive track report. In *Proceedings of the text REtrieval conference* (pp. 57–64).

Over P. (2001) The trec interactive track: an annotated bibliography. *Information Processing & Management, 37*(3):369–381

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: posters*, COLING '10 (pp. 997–1005). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ross, J., Zaldivar, A., Irani, L., Tomlinson, B., & Silberman, M. S. (2010). Who are the crowdworkers? shifting demographics in mechanical turk. In *Proceedings CHI 2010* (pp. 2863–2872).

Santos, R., Peng, J., Macdonald, C., & Ounis, I. (2010). Explicit search result diversification through subqueries. In: C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, & K. van

Rijsbergen (Eds.) *Advances in information retrieval, volume 5993 of lecture notes in computer science*. (pp. 87–99). UK: Springer.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*, (2nd edn.). Boston: Houghton Mifflin.

Voorhees, E. M. (2005). Trec: Improving information access through evaluation. *Bulletin of the American Society for Information Science and Technology, 32*(1), 16–21.

Voorhees, E. M., & Harman, D. (2005). *TREC: Experiment and evaluation in information retrieval digital libraries and electronic publishing*. Cambridge, MA: MIT Press.

Zuccon, G., Leelanupab, T., Whiting, S., Jose, E. Y. J., & Azzopardi, L. (2011a). Crowdsourcing interactions—Capturing query sessions through crowdsourcing. In B. Carterette, E. Kanoulas, P. Clough, & M. Sanderson (Eds.), *Proceedings of the workshop on information retrieval over query sessions at the European conference on information retrieval (ECIR)*. Dublin, Ireland, April 2011.

Zuccon, G., Leelanupab, T., Whiting, S., Jose, J., & Azzopardi, L. (2011b). Crowdsourcing interactions—A proposal for capturing user interactions through crowdsourcing. In M. Lease, V. Carvalho, & E. Yilmaz (Eds.), *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the 4th ACM international conference on web search and data mining (WSDM)* (pp. 35–38). Hong Kong, China, February 2011.